creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

공학석사학위논문

기계학습을 통한 한국주식시장 IPO의 상장일 주가 예측

Listing Day Price Prediction of Korean Stock Market IPO by Machine Learning

2023 년 2 월

서울대학교 대학원

산업공학과

이 종 우

기계학습을 통한 한국주식시장 IPO의 상장일 주가 예측

Listing Day Price Prediction of Korean Stock Market IPO by Machine Learning

지도교수 장 우 진

이 논문을 공학석사 학위논문으로 제출함

2022 년 12 월

서울대학교 대학원

산업공학과

이 종 우

이종우의 공학석사 학위논문을 인준함

2023 년 1 월

위 원 장	이 덕 주	(인)
부위원장	장 우 진	(인)
위 원	이 재 욱	(인)

초록

한국 증권시장에서의 공모(Initial Public Offering; 이하 IPO)는 비상장기업이 주식시장에 상장함을 의미한다. 공모주는 2020년 이후 많은 관심을 받고 있는 투자 방식으로 공모주의 저평가(Underpricing) 발행으로 인해 대부분 주가가 상승하는 모습을 보인다. 하지만 모든 IPO의 상장일 종가가 공모가보다 상승하지 않아 손해가 발생하고, 공모주 일정이 겹칠 경우 높은 수익을 가져올 수 있는 IPO에 더 많은 자산을 분배해야할 필요가 있다. 본 논문에서는 재무, 시장상황, IPO 데이터 변수 및 z-score 기관경쟁률과 같은 파생변수를 통해 상장일 IPO 종가를 분류 예측하는 기계학습 모델을 제안한다. 또한 설명가능한 인공지능(Explainable Artificial Intelligence; XAI) 방법 중 하나인 SHAP(Shapley Additive Explanation)를 통해 변수의 중요도 및 값에 따른 영향력을 분석했다. 실제 한국 증권시장의 IPO 데이터에서 파생변수를 활용한 인공신경망 모델이 92.6퍼센트의 등락 예측 정확도와 72.7 퍼센트의 150퍼센트 이상 상승한 IPO의 예측 정확도를 보였다. 또한 SHAP 값을 통해 가장 중요한 변수가 공모가, 기관경쟁률 그리고 의무보유확약비율임을 확인할 수 있었다. 공모가가 낮을수록, 기관경쟁률이 낮을수록 그리고 의무보유확약비율이 높을수록 모델이 IPO의 종가가 덜 하락하고 더 높게 상승한다고 예측하였고 이는 실제 사례 및 연구와 일치함을 확인하였다.

주요어: IPO, 기계학습, 설명가능한 인공지능, SHAP

학번: 2021-29800

목차

초록 **i**

목차 **iii**

표 목차 **iv**

그림 목차 **v**

제 1 장 서론 **1**

제 2 장 선행 연구 **5**

2.1 랜덤 포레스트 5

2.2 Extreme Gradient Boosting, XGBoost 5

2.3 로지스틱 회귀 6

2.4 서포트 벡터 머신 6

2.5 인공신경망 7

2.6 Shapley Additive Explanation, SHAP 7

2.7 기계학습을 통한 IPO 예측 선행 연구 8

제 3 장 방법론 **10**

3.1 종속 변수 10

3.2 주가지표 파생변수 생성 11

3.3 z-score 기관경쟁률 파생변수 생성 11

3.4 데이터 전처리 13

제 4 장 실험 **15**

4.1 데이터 15

4.2 사용변수 15

4.3 학습 및 평가 데이터 16

4.4 실험 진행 17

제 5 장 실험 결과 **19**

5.1 변수 및 모델에 따른 성능 19

5.2 변수 중요도 분석 22

제 6 장 결론 **30**

참고문헌 **34**

Abstract **38**

표 목차

표 3.1 공모가 및 총 공모주식수를 통한 주식투자지표 11

표 4.1 정보에 따른 변수 분류 16

표 4.2 사용 변수와 클래스에 따른 학습 및 훈련 데이터 수 16

표 4.3 변수 정의 18

표 5.1 변수 및 모델에 따른 정확도 19

표 5.2 변수 및 모델에 따른 혼동 행렬 20

표 5.3 변수 및 모델의 클래스별 정확도 20

표 5.4 기관경쟁률, 의무보유확약비율, 공모가만을 이용한 학습의 각 클래스별 정확도 27

표 5.5 기관경쟁률, 의무보유확약비율, 공모가만을 이용한 학습의 정확도 . 27

표 5.6 기관경쟁률 변수를 이용한 서포트 벡터 머신의 각 클래스별 중요변수 순위 29

표 5.7 z-score 기관경쟁률 변수를 이용한 인공신경망의 각 클래스별 중요변수 순위 29

표 5.8 기관경쟁률 변수 둘 모두를 이용한 인공신경망의 각 클래스별 중요변수 순위 29

그림 목차

그림 1.1 IPO 진행과정 1

그림 3.1 상장일 공모가 대비 종가 변화 도수분포표 10

그림 3.2 z-score기관경쟁률 생성 과정 11

그림 3.3 기관경쟁률 산점도 12

그림 3.4 z-score 기관경쟁률 산점도 13

그림 5.1 기관경쟁률 변수를 이용한 서포트 벡터 머신의 각 클래스별 SHAP 값 요약 그래프 23

그림 5.2 z-score 기관경쟁률을 이용한 인공신경망의 각 클래스별 SHAP 값 요약 그래프 24

그림 5.3 기관경쟁률 변수 둘 모두를 이용한 인공신경망의 각 클래스별 SHAP 값 요약 그래프 25

제 1 장 서론

한국 증권시장에서의 공모(Initial Public Offering; 이하 IPO)는 비상장기업이 일련의 과정을 거쳐 유가증권시장(코스피;KOSPI) 또는 코스닥(KOSDAQ) 시장에 상장함을 의미하며, 기업공개라고도 한다. 공모주는 IPO 과정에 참여한 기관투자자 및 일반청약자들에게 판매하는 주식을 의미한다.

공모주는 높은 수익률로 인해 2020년 이후 많은 관심을 받고 있는 투자방식이다. 공모주는 IPO의 흥행을 위해 본래 기업의 가치보다 저평가(Underpricing) 되어 발행되고, 이에 높은 수익률을 보이기 때문이다.[6] 한국 시장의 경우 2020년 코로나 19(COVID-19)에 의한 저금리 및 양적완화로 인해 증권시장의 호황이 찾아와[27] 사람들의 관심이 높은 기업들의 공모가 이루어졌다. 또한, 청약에 참여한 모든 참여자에게 균등하게 공모주를 배정하는 균등배정 제도가 2021년에 시작함으로 인해 개인투자자들의 관심도가 증가하였다. 한국 공모주의 상장일 공모가 대비 종가 기준 평균 수익률은 2021년도에 약 57.6퍼센트를 보였으며, 공모주 투자는 높은 수익률을 보이는 하나의 투자방법으로 각광받고 있다.

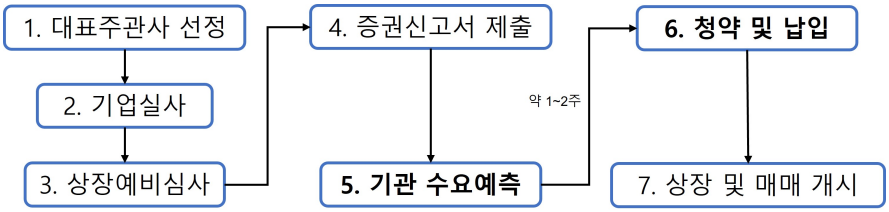


그림 1.1: IPO 진행과정

개인투자자는 일련의 과정을 거치는 공모에 청약 및 납입 단계에서 참여가 가능하며, 청약증거금을 제출하여 공모주를 배정받아 수익을 낼 수 있다. 공모주가 주식에

상장되기까지의 과정은 그림1.1와 같다. 기관 수요예측 단계에서 기관들이 IPO의 주관사가 제시한 가격 범위 내에서 입찰하고자한 가격을 통해 공모가가 결정된다. 이 과정을 통해 기관경쟁률, 의무보유확약비율과 같은 정보가 산출된다. 다음 단계인 청약 및 납입에서 일반청약자의 IPO 참여(청약)가 가능하며 공모주를 배정받기 위해 청약증거금을 필요로한다. 균등배정은 청약에 참여하기 위한 최소한의 청약증거금을 제출하며, 경쟁배정에 할당된 공모주를 배정받기 위해서는 최대한 많은 금액의 청약증거금이 필요하다. 이후 시장에 상장된 공모주의 시작가는 -10퍼센트에서 최대 +100퍼센트까지 변동하고 시작가에서 상장일의 종가는 주식상하한제도와 같이 ±30퍼센트 범위에서 변동이 가능하다. 공모주는 상장일에 최고 160퍼센트까지 상승이 가능하며 매우 큰 수익을 볼 수 있는 투자 방식이다.

개인투자자는 손해를 방지하고 더 큰 수익을 내기 위해 공모주의 상승 및 하락을 예측할 필요가 있다. 저평가된 공모주는 상장일에 대부분 상승하는 모습을 보이나 하락이 발생하는 경우가 있다. 하락한 IPO는 공모가와 주가의 하락에 의한 손해만이 아닌 청약증거금으로 인한 자금의 유동성 하락과 대출로 인한 이자손실이 발생한다. 또한 둘 이상의 IPO 일정이 겹치는 경우 한정된 자본을 통해 어느 IPO에 자산을 더 많이 분배할 것인지에 대한 문제가 발생한다. 기관투자자에 비해 일반청약자들은 가지고 있는 정보의 양이 적지만, 기관 수요예측 단계 이후 일반 청약 전까지 시간의 차이를 통해 기관이 IPO에 참여한 정보를 통해 정보의 불균형을 소폭 해소할 수 있다. 일반청약자는 청약에 참여 결정 및 상장 후 주식을 투자할지에 대해 기관 수요예측 단계가 끝난 이후 판단이 가능한 기회가 존재한다. 따라서 하락하는 IPO를 회피하고 더욱 크게 상승할 IPO를 예측하는 방법이 일반 청약자들의 청약 참여 및 투자 결정에 있어 필요하다.

공모주의 등락을 적정가 회귀 및 분류를 통해 예측하는 연구는 지속적으로 이루어지고 있다. 해외 공모주 연구 사례로는 미국 증권시장에서 상장일 지수수익률 대비

공모주의 수익률을 IPO 관련 변수를 통해 랜덤포레스트 모델을 통한 회귀분석 연구[24]와 터키 증권시장 IPO에서 상장일 공모가 대비 종가 수익률이 중앙값보다 큰 경우를 서포트 벡터 머신(Support Vector Machine; SVM)과 랜덤 포레스트(Random Forest) 모델을 사용하여 예측한 연구[4]가 있다. 해외 시장의 공모 예측 연구는 주식 상하한제 유무 및 정도의 차이와 균등배정과 같은 제도 차이에 의해 국내 IPO 예측 연구에 그대로 반영하기 어려운 한계가 있다. 국내 연구 사례로는 상장 후 20일 평균 종가를 동일 산업군 기업들의 재무 정보를 통해 파생된 변수들로 학습한 인공신경망 회귀분석 연구[15], 비재무 정보를 통해 로지스틱 회귀(Logistic Regression), 의사결정 나무(Decision Tree), 서포트 벡터 머신 모델 등을 학습하여 상장일 가격 등락 예측[30]과 인공신경망을 활용한 공모주 투자 및 기준 수익률 달성 여부 예측 모델 연구[31]가 있다. 국내 연구는 다양한 변수들과 학습 모델들을 사용하였으나 비재무, 재무 그리고 시장 상황와 같은 모든 변수와 이로 인해 파생된 변수를 활용한 연구는 적었다. 또한 종가의 예측에 관한 연구는 많으나 어떠한 변수로 인하여 공모가 대비 종가가 하락 또는 상승하였는지, 중요한 영향을 끼친 변수는 무엇인지에 대해 분석하지 않은 한계가 존재한다.

이에 본 논문에서는 재무, 시장상황, IPO 데이터와 데이터를 활용하여 생성된 파생변수를 통해 상장일 공모주의 등락을 예측한다. 또한 SHAP(Shapley Additive Explanation) 기법을 사용하여 변수들의 중요도 및 변수의 값에 따른 예측값에 영향을 파악하였다. 이에 대한 결과를 확인하기 위해 증권 시장 IPO 데이터를 통하여 다양한 기계학습 모델을 학습하였다. 인공신경망 기반의 학습 모델이 상징일 종가 등락 예측 정확도 92.6퍼센트와 150퍼센트 이상 상승율 IPO 예측 정확도 72.7 퍼센트를 보였다. 또한 SHAP을 통한 결과가 실제 IPO에 영향을 끼치는 변수들의 사례 및 연구와 일치함을 확인하였다.

본 논문의 구성은 다음과 같다. 제 2장에서는 학습 모델으로 사용한 랜덤 포레스트, XGBoost(Extreme Gradient Boosting), 로지스틱 회귀, 서포트 벡터 머신 그리고 인공신경망(Artificial Neural Network; ANN)과 XAI(Explainable Artificial Intelligence) 방법인 SHAP에 대하여 소개한다. 제 3장은 실험에 사용한 데이터와 상장일 공모가 대비 종가의 상승폭을 통해 생성한 종속 변수에 대해 살펴본다. 제 4장에서는 주가지표, z-score 기관경쟁률 파생변수 생성 및 데이터 전처리과정의 방법론에 대한 내용을 설명한다. 제 5장에서는 사용 변수, 수집한 데이터 그리고 실험 진행 과정을 소개한다. 제 6장은 앞 장의 실험 진행 과정을 적용하여 변수 및 모델에 따른 예측 성능 실험 결과를 소개하고 SHAP을 통해 도출한 변수 분석 결과가 실제 IPO 사례 및 연구와 일치함을 설명한다. 제 6장은 본 연구의 결과 요약과 한계점 및 후추 과제에 대하여 논의한다.

제 2 장 선행 연구

2.1 랜덤 포레스트

랜덤 포레스트는 의사결정나무의 앙상블(Ensemble) 기법 중 하나이다. [7] k번째 의사결정나무에서 생성된 각 벡터 Θ_k는 다른 의사결정나무의 벡터와 독립이지만 같은 분포를 가진다. 생성된 벡터 Θ_k와 학습 데이터를 통해 입력값 $\mathbf{x}$를 분류하는 분류기 $h(\mathbf{x}, \Theta_k)$를 구성할 수 있다. 각 의사결정나무의 분류기를 통해 나온 결과를 배깅(Bootstrap Aggregation; Bagging)을 통해 결정한다.

2.2 Extreme Gradient Boosting, XGBoost

XGBoost는 그래디언트 부스팅(Gradient Boosting)의 단점을 해결한 오픈 소스 라이브러리(Open Source Library)이다.[8] 부스팅(Boosting)은 순차적으로 여러 개의 약한 분류기를 결합하여 강한 분류기를 생성하는 앙상블 기법이다. 부스팅 방법 중 하나인 그래디언트 부스팅은 분류기를 만들 때, 현재 약한 분류기의 손실 함수를 최소화하는 방향의 기울기(Gradient)를 통해 새로운 분류기를 학습하는 방법이다.[12] 하지만 과적합이 발생하고 병렬 학습이 어려운 단점이 존재한다. 그래디언트 부스팅 기반으로 구현된 XGBoost가 이러한 문제점을 해결하기 위해 제안되었다. XGBoost는 그래디언트 부스팅 대비 과적합 규제(Regularization)가 가능하고 학습 및 분류 속도가 빠르다는 장점이 있다.

2.3 로지스틱 회귀

로지스틱 회귀란 각 변수들의 선형 결합을 통해 분류하는 통계학습 기법이다.[17] 선형 회귀 분석과는 달리 종속 변수가 범주형 데이터일 경우에 사용한다. 독립 변수 $\mathbf{x}$가 주어졌을 때 $\mathbf{x}$가 분류되는 결정 경계를 식2.1와 같이 표현할 수 있다. 결정 경계 식에 로그(log)를 양변에 취하면 식2.2와 같다. 로짓(Logit)이 선형이라고 가정할 때 식 2.3와 같이 표현이 가능하며 $\mathbf{x}$가 종속 변수 $\mathbf{Y}$의 1에 속할 확률 값인 로지스틱 함수 2.4를 구할 수 있다.

$$\{\mathbf{x} : \Pr(Y = 1|\mathbf{X} = \mathbf{x}) = 0.5\} \tag{2.1}$$

$$\{\mathbf{x} : log(\Pr(Y = 1|\mathbf{X} = \mathbf{x}/\Pr(Y = 0|\mathbf{X} = \mathbf{x}) = 0\} \tag{2.2}$$

$$log(\Pr(Y = 1|\mathbf{X} = \mathbf{x}/\Pr(Y = 0|\mathbf{X} = \mathbf{x}) = \beta_0 + \mathbf{x}'\beta \tag{2.3}$$

$$\Pr(y = 1|\mathbf{X} = \mathbf{x}) = \frac{\exp(\beta_0 + \mathbf{x}'\beta)}{1 + \exp(\beta_0 + \mathbf{x}'\beta)} \tag{2.4}$$

2.4 서포트 벡터 머신

서포트 벡터 머신은 데이터를 분류하는 가장 큰 마진(Margin)을 가진 초평면을 선택하는 학습 모델이다. 초평면에 의해 구분된 데이터의 각 서포트 벡터를 지나는 초평면 사이의 거리를 마진이라고 한다. 마진은 $\frac{2}{||w||}$이며 w는 초평면의 법선 벡터(Normal Vector)이다. 하지만 모든 데이터에서 자료들을 완벽히 분리하는 초평면은 존재하지 않을 수 있으며 이러한 문제를 해결하기 위해 소프트 마진(Soft Margin) SVM이 제안되었다.[10] 목적함수에 여유 변수(Slack variable) ξ를 부여함으로써 마진을 크게 하되 에러에 대한 페널티(Penalty)를 작게 하는 방향으로 초평면을 학습한다. 소프트 마진 SVM은 식2.5와 같은 최적화 문제로 표현된다.

$$\text{minimize} \quad ||w||_2^2 + \gamma \sum_{i \in X} \xi_i \tag{2.5}$$

$$\text{subject to} \quad y_i(x_i^T w - b) \geq 1 - \xi_i \qquad i \in X, \tag{2.6}$$

$$\xi_i \geq 0 \qquad i \in X. \tag{2.7}$$

2.5 인공신경망

인공신경망(Artificial Neural Network, ANN)은 뇌의 신경망 구조에서 영감을 얻은 통계학적 학습 알고리즘으로 입력층, 은닉층과 출력층으로 이루어져 있다. 각 층은 하나 혹은 다수의 노드(Node)로 구성되어 있으며 노드들을 연결하는 가중치의 학습을 통해 모델을 최적화 한다. 학습된 가중치를 반영한 값을 활성함수(Activation function)을 통해 출력하여 다음 노드로 전달하며 가중치는 역전파 알고리즘(Backpropagation algorithm)을 통해 학습한다. 인공신경망을 통해 비선형 데이터에서 원하는 함수를 추론할 수 있다.[14]

2.6 Shapley Additive Explanation, SHAP

SHAP 값은 기존 학습 모델에 대한 조건부 기대 함수의 섀플리 값(Shapley Value)으로 계산된 변수의 중요도 값이다.[21] 섀플리 값이란 어떠한 과제에 참여한 참가자들의 공헌도를 나누는 이론이다. [13] 기계학습에서 모델의 예측값을 과제로, 변수들의 영향력을 참가자들의 공헌도로 볼 수 있으며, 섀플리 값을 통해 각 변수가 모델의 예측에 끼치는 영향력을 계산할 수 있다.[19] SHAP 값은 각 변수들이 변할 때 모델이 예측한 값이 얼마만큼 변화하는 지를 계산하며, 식 2.8와 같이 SHAP 값 ϕ_i을 얻고자 하는 변수

i를 포함하는 모델(F)과 포함하지 않는 모델($S \subseteq F$)의 예측을 비교하여 구한다.

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(F - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)] \tag{2.8}$$

SHAP을 통해 각 변수의 영향력을 알 수 있으며, 변수의 입력 값이 변할 때 결과가 어떻게 변하는지 해석이 가능하다. SHAP 값의 정확한 계산은 상당한 자원이 요구되며 커널(Kernel) SHAP 방법을 통해 모든 학습 모델에서 근사값을 계산할 수 있다. 각 학습모델에 적합한 SHAP 모델을 사용하여 커널 SHAP보다 빠르게 근사값의 계산이 가능하다. 선형 회귀 기반 모델의 경우 리니어(Linear) SHAP을, 의사결정나무 기반 앙상블 모델의 경우 트리(Tree) SHAP[20]을 그리고 인공신경망 기반 모델의 경우 DeepLIFT[25]와 섀플리 값을 조합한 딥(Deep) SHAP을 통해 근사값을 구한다.

2.7 기계학습을 통한 IPO 예측 선행 연구

IPO를 예측하는 연구는 크게 분류와 회귀 지도학습으로 나눌 수 있으며 국내외로 다양하게 연구가 진행되었다. 해외 연구 사례로는 미국 증권시장에서 상장일 지수수익률 대비 공모주의 상장일 종가 수익률을 IPO 관련 변수들을 이용하여 랜덤포레스트 모델을 통해 회귀분석을 한 연구가 있다.[24] S&P500, AMEX 그리고 NASDAQ에 상장된 866개의 기업들에 대한 연구를 수행하였다. 랜덤포레스트 모델의 학습에는 공모가 밴드 크기, 공모가, 기술주 여부 그리고 공모규모 등 IPO 관련 정보 총 6개의 변수를 사용하였다. 다른 해외 연구로는 터키의 주식시장인 Borsa Istanbul에 2005년부터 2013년까지 상장된 기업들의 종가 수익률을 회귀분석 하여 중심경향치 보다 큰 경우를 1로 작은 경우를 0으로 예측하는 실험을 하였다.[4] 사용 변수로는 기업 연식, 매출 그리고

자산과 같은 기업 정보, 21일 주식 인덱스 수익률과 같은 주식 시장 상황 정보 그리고 공모가와 같은 IPO 관련정보 모두를 학습에 사용하였다. 의사결정나무와 서포트 벡터 머신을 10겹 교차검증을 통해 학습하였다. 해외 IPO 연구는 학습에 사용한 변수와 종속 변수에 대한 정보를 알 수 있으나 국내 IPO와 제도의 차이에 의해 온전히 반영할 수 없다는 한계가 존재한다.

국내 연구 사례로는 인공신경망을 통해 코스닥 상장 기업들의 IPO 후 20일 종가를 예측하는 연구가 있다. [15] 동일한 산업군에 속한 기업들의 재무 정보를 통해 IPO 기업의 자산을 주가수익비율과 같은 주가지표를 통하여 인공신경망 학습을 진행하였다. 공모규모, 신주모집 비율 그리고 의무보유확약 비율과 같은 비재무 정보만을 통해 로지스틱 회귀분석, 의사결정나무, 인공신경망, 사례기반추론 그리고 서포트 벡터 머신을 학습시켜 상장일 종가의 등락을 예측한 연구[30]가 있다. 입력 변수를 일대일 t검정, 로지스틱 회귀분석 그리고 의사결정 나무를 통해 결정하여 의사결정나무, 로지스틱 회귀분석 및 인공신경망 모델 등을 통해 기준 수익률 20%를 초과하거나 하지 못하는 IPO를 예측하는 연구 또한 이루어졌다.[31] 연구 결과 상장시기, 기관경쟁률, 공모액, 밴드상하단 가격과 같은 IPO관련 정보 7개를 사용하여 학습한 인공신경망 모델이 가장 좋은 성능을 보였다. 국내 연구들은 본 연구와 시장에 따른 제도가 일치하여 쉽게 적용이 가능한 장점이 있으나 기업 관련, 시장 상황 그리고 공모 관련 정보 모두를 사용하여 활용한 연구가 적었다. 또한 예측하고자 하는 종가에 관한 연구는 많으나 모델의 예측이 어떠한 변수를 중요하게 받아들였으며, 변수의 크기에 따라 어떻게 예측하였는지에 관해 분석하지 않은 한계가 있다.

제 3 장 방법론

3.1 종속 변수

그림3.1은 상장일의 공모가 대비 종가 수익률 도수분포표로 수익률을 통해 IPO 데이터를 3가지 클래스(Class)로 나누어 학습하였다. 첫번째 클래스는 공모가 보다 종가가 떨어져 손해를 보는 IPO이며 두번째 클래스는 수익률이 0퍼센트 이상 150퍼센트 미만인 IPO, 그리고 마지막 클래스는 수익률이 150퍼센트 이상인 IPO이다. 이는 한국 주식시장의 제도에 의해 공모가 대비 시작가는 최소 -10퍼센트에서 최대 100퍼센트까지 등락이 가능하며, 시작가 대비 공모가는 ±30퍼센트임에 기인한다. 즉, IPO 주식의 상장일 등락은 최소 -37퍼센트에서 최대 160퍼센트이다. 본 논문에서는 첫번째 클래스를 0, 두번째 클래스를 1 그리고 마지막 클래스를 2로 라벨링(Labeling)하여 학습을 진행하였다.

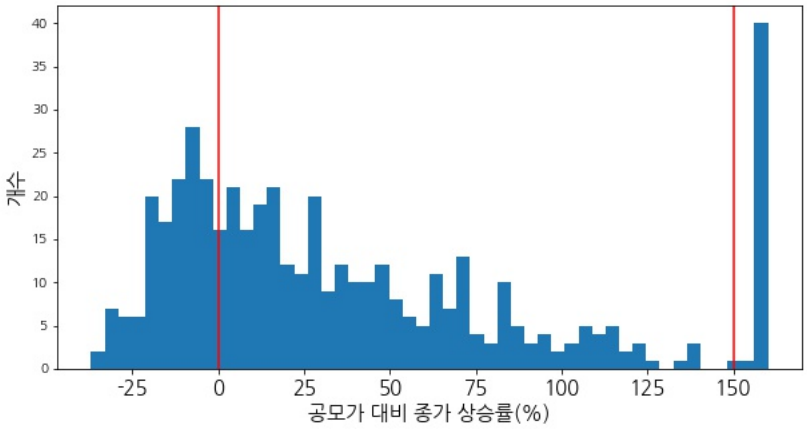


그림 3.1: 상장일 공모가 대비 종가 변화 도수분포표

3.2 주가지표 파생변수 생성

주식 투자에 있어 사용되는 지표인 자기자본이익률(Return On Equity, ROE), 주당순이익(Earnings Per Share, EPS), 주가수익비율(Price Earnings Ratio, PER)[5], 주당순자산(Book-value Per Share, BPS) 그리고 주가순자산비율(Price Book-value Ratio, PBR)[11]을 파생변수로서 생성하였다. 해당 변수들을 계산할 때 주가로는 공모가를 총 주식수로는 총 공모주식수를 이용하였으며 표3.1와 같이 계산된다. 주가지표 파생변수들을 통하여 공모주가 IPO를 진행하는 기업의 가치에 비해 고평가되었는지 여부를 학습할 수 있다.

자기자본이익률	$\text{ROE} = \frac{\text{최신분기순이익}}{\text{자본}} \times 100$
주당순이익	$\text{EPS} = \frac{\text{최신분기순이익}}{\text{총 공모주식수}}$
주가수익비율	$\text{PER} = \frac{\text{공모가}}{\text{주당순이익}}$
주당순자산	$\text{BPS} = \frac{\text{자산}}{\text{총 공모주식수}}$
주가순자산비율	$\text{PBR} = \frac{\text{공모가}}{\text{주당순자산}}$

표 3.1: 공모가 및 총 공모주식수를 통한 주식투자지표

3.3 z-score 기관경쟁률 파생변수 생성

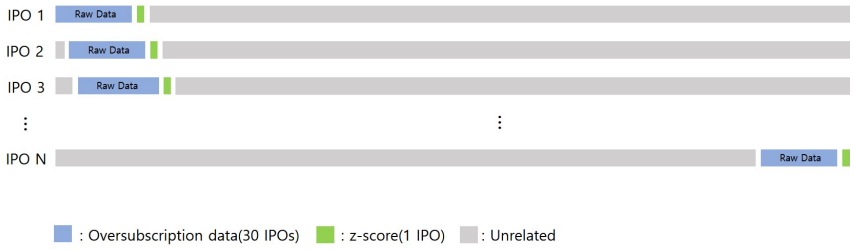


그림 3.2: z-score기관경쟁률 생성 과정

본 논문에서는 그림3.2와 같은 롤링(Rolling)을 과정을 통해 z-score 기관경쟁률 변수를 생성하여 학습에 이용하였다. 2022년으로 갈수록 기관경쟁률이 증가하며 양

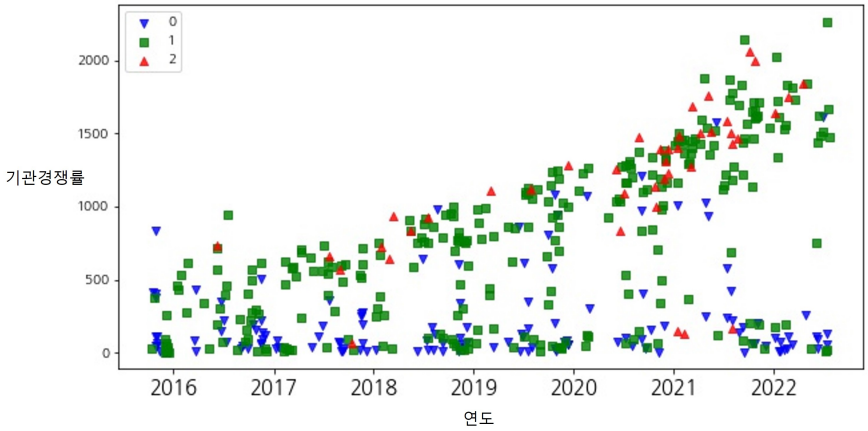


그림 3.3: 기관경쟁률 산점도

극화되는 현상을 그림3.3에서 확인할 수 있다. 기관경쟁률을 그대로 이용할 경우 IPO 시장의 과열성을 반영하지 못하여 기관경쟁률을 통한 학습이 정확히 되지 않을 수 있다. 2022년의 기관경쟁률이 2017년의 기관경쟁률보다 절대적인 수치가 크다고하여 공모가 대비 시가의 상승폭이 더 높지 않은 경우가 있음을 그림3.3를 통해 확인할 수 있다. 이러한 기관경쟁률의 절대적인 크기가 상장일 종가 예측의 학습에 시장의 과열을 고려하지 않고 반영되는 문제를 해소하기 위하여 z-score 파생변수를 생성하였다. 학습 및 평가하고자하는 IPO에서 가장 최근에 종료된 30개의 IPO의 기관경쟁률과 현 IPO의 기관경쟁률을 통해 z-score 기관경쟁률 파생변수를 생성하였으며 계산방법은 식3.1과 같다. z-score 기관경쟁률 계산에서 30개가 사용된 이유는 중심극한정리가 성립하기 위한 최소 표본의 수를 맞추기 위함이다.[22] 그림 3.4는 z-score 기관경쟁률을 산점도로 나타낸 것이다. 산점도를 통해 상장일 종가의 하락한 IPO를 의미하는 라벨 0이 대부분 z-score -1에, 종가가 크게 상승한 라벨 2가 1 주변에 위치한 것을 볼 수 있다. 또한 연도와 관계없이 z-score 기관경쟁률은 대부분 2와 -2 사이에 분포하는 것을 확인할 수 있다.

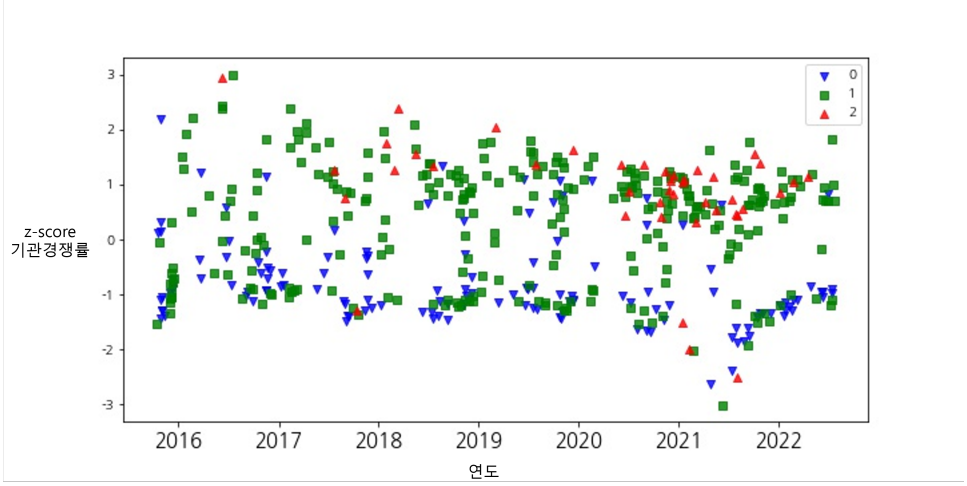


그림 3.4: z-score 기관경쟁률 산점도

$$(\text{z-score 기관경쟁률}) = \frac{(\text{기관경쟁률}) - (\text{30개 IPO 기관경쟁률의 평균})}{(\text{30개 IPO 기관경쟁률의 표준편차})} \tag{3.1}$$

3.4 데이터 전처리

데이터 전처리는 z-score 정규화와 랜덤 오버샘플링을 적용하였다. 로지스틱 회귀 분석, 서포트 벡터 머신 그리고 인공신경망 학습의 경우 공모가 밴드위치와 z-score 기관경쟁률을 제외한 수치형 자료에 z-score 정규화를 통해 전처리 과정을 진행하였다. 공모가 밴드 위치는 공모가를 공모가 하단과 공모가 상단에 대하여 최소-최대 정규화를 진행한 변수이며 z-score 기관경쟁률은 z-score 정규화를 통해 나온 파생변수이므로 z-score 정규화 전처리 과정에서 제외하였다. 또한 데이터 불균형을 해소하기 위해 랜덤 오버샘플링을 훈련 데이터에 적용하였다. 랜덤 오버샘플링(Random Oversampling)은 특정 클래스의 데이터가 부족할 때 그 크기를 늘림으로써 데이터 불균형에 의한 과적합을 해소할 수 있는 방법이다. 언더샘플링을 적용하지 않은 이유는 기존 학습 데이터의

수가 400개 미만으로 언더샘플링으로 인해 데이터의 크기가 더 줄어들 경우 모델의 학습이 잘 진행되지 않을 수 있기 때문이다. 랜덤 오버샘플링은 클래스의 분포가 불균형인 전체 데이터가 S이고 수가 적은 클래스의 데이터가 S_{min}일 때, S_{min}에서 정해진 E만큼 데이터를 임의로 추출하고 복제하여 S에 추가하는 방식으로 진행된다. S에서의 소수 클래스의 크기를 증가시켜 데이터 분포의 불균형을 해결하며 실행과정이 단순하다는 장점이 있다.[16]랜덤 오버샘플링을 통해 데이터의 수가 가장 많은 클래스 1에 클래스 0과 2의 데이터 크기를 동일하게 맞추어 데이터 불균형을 해소하고자 하였다.

제 4 장 실험

4.1 데이터

본 논문에서는 유가증권시장과 코스닥에 상장된 기업들의 IPO 데이터를 사용한다. 한국 주식시장의 가격제한폭이 기존 ±15퍼센트에서 현재 적용중인 폭인 ±30퍼센트로 변경된 2015년 6월 이후, 2022년 7월까지 약 7년의 데이터를 수집하였다. 상장하였으나 2022년 7월 전 상장폐지된 기업들은 특정 정보의 누락으로 인해 학습 데이터에서 제외하였다. 그리고 일반공모주와 성격이 다른 스팩(Special Purpose Acquisition Company; SPAC)주 [1] 와 리츠(Real Estate Investment Trusts; REITs)주 [2] 를 학습 데이터에서 제외하였다. 스팩주와 리츠주는 일반공모주의 성격적 차이 외, 제도적으로 IPO 과정에서 차이가 존재하여 학습데이터에서 제외하였다. 최종적으로 상장폐지, 스팩 및 리츠주를 제외한 총 499개의 IPO 기업을 학습 및 평가 데이터로 사용하였다.

4.2 사용변수

본 논문에서는 정보의 출처에 따라 3종류로 구분되는 기본변수와 이로부터 생성된 파생변수들을 통해 모델을 학습하였다. 정보의 출처에 따른 기본변수는 표4.1에서 확인이 가능하며 변수에 대한 정의 및 계산 방법은 표4.3와 같다.

또한 수집된 데이터를 이용하여 생성된 파생변수들은 자기자본이익률, 주당순이익, 주가수익비율, 주당순자산, 주가순자산비율과 z-score 기관경쟁률이다.

[1] 다른 기업과 합병하는 것을 사업 목적으로 하여 주식을 발행하는 법인이다. 일반투자자들은 향후 합병이 완료된 후 발생할 수 있는 투자이익을 기대하고 스팩주의 IPO에 참여한다.

[2] 리츠란 '자산을 부동산에 투자하여 운용하는 것을 주된 목적으로 부동산투자회사법에 따라 설립된 회사'로 부동산투자회사법 제2조 제1호에 정의되어 있다. 부동산투자회사가 발행하는 주식이 리츠주로써 회사가 부동산에 투자, 운용하여 얻은 투자수익을 주주들에게 배당한다.

회계 및 기업 정보	주식 시장 동향 정보	IPO 관련 정보
자산 부채 자본 최신분기매출 최신분기순이익 해당 산업군	수요조사 X일전 시장 지수 상승율 (X=15,30,60,90,150,300) 수요조사 Y일전 산업군 지수 상승율 (Y=15,30,60,90,150,300)	총 공모주식수 기관경쟁률 신주모집비율 공모가 밴드위치 공모가 공모규모 의무보유확약 비율 균등분배여부

표 4.1: 정보에 따른 변수 분류

4.3 학습 및 평가 데이터

Variable	Train				Test			
	Class			합계	Class			합계
	0	1	2		0	1	2	
기관경쟁률	117	247	34	398	30	59	12	101
z-score기관경쟁률 둘 모두 사용	112	231	31	374	27	57	11	95

표 4.2: 사용 변수와 클래스에 따른 학습 및 훈련 데이터 수

학습 및 평가데이터는 전체 데이터를 4대 1로 나누었으며 롤링을 통해 생성된 파생변수에 의해 실험에 따라 데이터의 크기가 다르다. 또한 각 클래스에 해당되는 IPO의 수가 그림3.1에서와 같이 큰 차이가 있으며 그 수는 표4.2과 같다. 실험에 사용한 데이터가 금융데이터임을 고려하여 학습 및 시험 데이터를 무작위로 추출하는 것이 아닌 시간을 기준으로 나누었으며 시기는 각각 2021년 4월과 5월이다. 학습 데이터는 기관경쟁률 및 z-score 기관경쟁률을 이용한 실험에 따라 각 398개와 374개 이며 평가데이터는 각 101개와 95개로 약 4대 1의 비율로 나누었다.

4.4 실험 진행

랜덤 포레스트, 로지스틱 회귀, 서포트 벡터 머신, XGBoost 그리고 인공신경망을 학습에 사용하는 기관경쟁률의 조합을 기관경쟁률, z-score 기관경쟁률 그리고 둘 모두를 사용하는 방식으로 총 3번에 걸쳐 실험을 진행하였다. 기관경쟁률 외 나머지 30개의 변수는 모두 동일하게 학습에 이용하였다. 기관경쟁률을 사용하여 학습하는 실험의 학습데이터의 수는 오버샘플링을 통해 741개를 사용하였으며, z-score 기관경쟁률 및 둘 모두를 사용하는 실험의 학습데이터의 수는 693개이다. 이는 오버샘플링을 통해 학습데이터 각 클래스에 해당하는 데이터의 크기를 그 수가 가장 많은 클래스 1에 맞추었기 때문이다.

학습 모델의 결과는 혼동 행렬(Confusion Matrix)을 통해 확인하였으며 평가지표는 정확도(Accuracy)를 사용하였다. 혼동 행렬의 경우 각 클래스에 대한 예측이 얼마나 정확했는지를 확인 및 평가하기 위해 사용하였으며 정확도는 각 클래스 별 평가 데이터의 수가 달라 전체적인 성능을 확인하기 위해 사용하였다. 본 논문에서는 전체 평가 데이터에 대한 정확도가 70% 이상, 혼동 행렬에서 각 클래스에 대한 예측 정확도가 60%이상의 수준을 보인 모델에 대하여 SHAP 값의 계산 및 변수 중요도 분석을 진행하였다. 이는 성능이 좋은 모델은 클래스의 구분을 잘함을 의미하고, 각 클래스들을 구분할 때 중요하게 작용하는 변수들을 SHAP 값을 통해 정확하게 제시할 확률이 더 높기 때문이다. 각 학습 모델에 최적화된 알고리즘을 통해 SHAP 값의 근사값을 계산하였다. 의사결정나무 기반 모델인 랜덤 포레스트(RF)와 XGBoost(XGB)는 트리 SHAP 알고리즘을[20], 선형회귀 분석 기반 모델인 로지스틱 회귀 분석(LR)은 리니어 SHAP을, 그리고 인공신경망 기반 모델(ANN)은 딥 SHAP 알고리즘을 사용하여 SHAP 값을 계산하였다. 서포트 벡터 머신(SVM)은 적합한 알고리즘이 존재하지 않아 모든 알고리즘의 SHAP 값의 근사가 가능한[21] 커널 SHAP 방법을 사용하였다.

변수	정의
자산	상장 시 IPO를 위한 투자설명서의 재무에 관한 사항에 기록된 최신 분기 자산
부채	상장 시 IPO를 위한 투자설명서의 재무에 관한 사항에 기록된 최신 분기 부채
자본	상장 시 IPO를 위한 투자설명서의 재무에 관한 사항에 기록된 최신 분기 자본
최신분기매출	상장 시 IPO를 위한 투자설명서의 재무에 관한 사항에 기록된 최신 분기 매출
최신분기순이익	상장 시 IPO를 위한 투자설명서의 재무에 관한 사항에 기록된 최신 분기 순이익
해당 산업군	IPO 기업의 한국증권거래소 지수산업분류에 해당되는 산업분류
수요조사 X일전 시장 지수 상승율	수요조사일과 X일전 IPO 상장 시장 지수의 변화율
수요조사 X일전 산업군 지수 상승율	수요조사일과 X일전 해당되는 산업 지수의 변화율
총 공모주식수	상장되는 주식의 수
기관경쟁률	기관청약 단계에서 (기업들이 청약한 주식 수)/(총 공모주식 수)
신주모집비율	(총 공모주식수)/(기업 전체 주식의 수)
공모가	기관투자자들에 의해 결정된 상장 주식 가격
공모가 밴드위치	주관사가 제시한 공모가 상단과 하단에서 공모가의 상대적 위치 ({(공모가)-(공모가하단)}/{(공모가상단)-(공모가하단)}
공모규모	공모가와 총 공모주식수로 계산된 시가총액 (총 공모주식수)*(공모가)
의무보유확약 비율	기관이 일정기간 의무보유하는 주식의 전체에서의 비율
균등배정여부	균등배정 제도가 시행되었는지 여부(2021년 이후 모든 청약에서 시행)

표 4.3: 변수 정의

제 5 장 실험 결과

5.1 변수 및 모델에 따른 성능

	RF	XGB	LR	SVM	ANN
기관경쟁률	69.3%	70.3%	65.3%	72.3%	74.3%
z-score	65.3%	71.6%	60.0%	64.2%	72.6%
Use both	74.7%	70.5%	61.1%	67.3%	71.6%

표 5.1: 변수 및 모델에 따른 정확도

학습 모델 성능은 표5.1와 같다. 두 기관경쟁률 변수를 모두 사용한 랜덤 포레스트 모델이 74.7 퍼센트의 정확도로 가장 좋은 성능을 보여주었으며 기관경쟁률을 이용한 인공신경망과 z-score 기관경쟁률을 이용한 인공신경망 그리고 기관경쟁률을 사용한 서포트 벡터 머신이 그 뒤를 따랐다. 평가 데이터의 약 60 퍼센트가 클래스 1 데이터이므로 모델이 모두 클래스 1로 예측한다면 최소 60 퍼센트의 정확도를 보여주게 된다. 이에 본 논문에서는 전체 예측에 있어 정확도만이 아닌 각 클래스에 대한 정확도를 확인하고자 각 클래스에 대한 혼동 행렬 결과를 확인하였으며 이는 표5.2 와 같다. 좌측이 실제 클래스이며 위쪽은 예측 클래스를 의미한다.

표5.3는 변수 및 모델에 따른 각 클래스별 정확도를 정리한 것이다. 기관경쟁률과 나머지 변수들을 사용하여 서포트 벡터 머신을 학습한 결과 전체 평가 데이터에서 72.3 퍼센트의 정확도를 보여주었으며 각 클래스별 정확도는 80 퍼센트, 69.5 퍼센트 그리고 66.7 퍼센트이다. z-score 기관경쟁률과 기타 정보들을 사용하여 배치 크기 16, 활성함수로는 ReLU(Rectified Linear Unit) 그리고 최적화 알고리즘으로는 Adam(Adaptive Moment Estimation)[18]을 사용한 각 32개와 2개의 노드로 구성된 은닉층 구조를 지닌 인공신경망을 학습한 결과 전체 평가 데이터에서 72.6 퍼센트의 정확도와 각 클래스에

		RF			XGB			LR			SVM			ANN		
		0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
기관경쟁률	0	25	2	3	18	12	0	15	15	0	24	6	0	27	3	0
	1	7	39	13	9	50	0	5	42	12	11	41	7	10	40	8
	2	0	6	6	0	9	3	0	3	9	1	3	8	0	5	7
z-score	0	24	2	1	21	6	0	11	16	0	20	7	0	26	0	1
	1	15	33	9	13	44	0	7	41	9	14	36	7	14	36	7
	2	0	6	5	0	8	3	1	5	5	1	5	5	2	2	7
Use both	0	21	6	0	21	6	0	13	14	0	23	4	0	25	1	1
	1	9	44	4	13	42	2	7	37	13	12	34	11	11	35	11
	2	0	5	6	1	6	4	1	2	8	1	3	7	1	2	8

표 5.2: 변수 및 모델에 따른 혼동 행렬

		RF	XGB	LR	SVM	ANN
기관경쟁률	0	83.7%	60.0%	50.0%	80.0%	90.0%
	1	61.1%	84.7%	71.2%	69.5%	69.5%
	2	50.0%	25.0%	75.0%	66.7%	58.3%
z-score	0	88.9%	77.8%	40.7%	74.1%	96.3%
	1	57.9%	77.2%	71.9%	63.2%	63.2%
	2	45.5%	27.3%	45.5%	45.5%	63.6%
Use both	0	77.8%	77.8%	48.1%	85.2%	92.6%
	1	77.2%	73.7%	64.9%	59.6%	61.4%
	2	54.5%	36.4%	72.7%	63.6%	72.7%

표 5.3: 변수 및 모델의 클래스별 정확도

서 96.3, 63.2 그리고 63.6 퍼센트의 정확도를 보여주었다. 두 기관경쟁률 모두와 30개 변수들을 사용하여 배치 크기 16, 활성함수로는 ReLU 그리고 최적화 알고리즘으로는 Adam을 사용한 각 16개와 2개의 노드로 구성된 은닉층 구조를 지닌 인공신경망을 학습시킨 결과 평가 데이터 정확도는 71.6 퍼센트이며 각 클래스 별 정확도는 92.6, 61.4 와 72.7 퍼센트이다. 기준을 만족한 3개의 모델 중 클래스 0과 2에서 종합적으로 높은 정확도를 보인 두 기관경쟁률 모두 이용한 인공신경망 모델을 최종 상장일 종가 등락 예측 모델로 결정하였다.

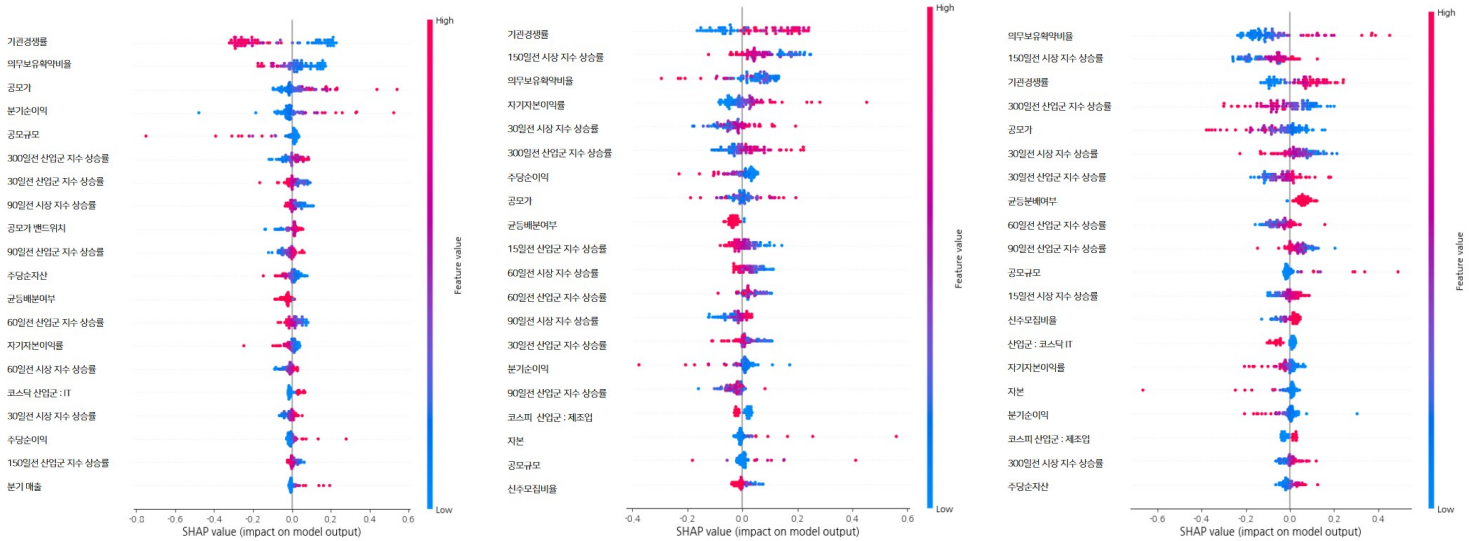
인공신경망 모델에서는 클래스 0을 90퍼센트 이상 올바르게 구분하는 결과를 보였으며 이는 종가가 공모가 대비 하락한 IPO를 90퍼센트 이상 맞추었음을 의미한다. 서포트 벡터 머신 모델은 클래스 0 구분 정확도가 80퍼센트로 인공신경망 모델 둘 모두에 비하여 공모가 대비 종가가 하락한 IPO의 예측 성능이 낮았다. 종가가 상승한 IPO 데이터에서 상승폭이 150 퍼센트 이상인 종목과 아닌 종목을 구분하는 것에 있어 약 60 퍼센트의 정확도를 보여주었다. 이는 학습 모델이 IPO 주가의 등락은 주어진 변수를 통해 잘 예측하나 상승폭을 구분함에 있어 상대적으로 학습이 잘 이루어지지 않았음을 알 수 있다.

인공신경망 모델을 z-score 기관경쟁률 혹은 두 기관경쟁률 모두를 이용하여 학습한 경우 기관경쟁률을 통해 학습한 클래스 0과 2의 학습정확도가 증가하였다. 기관경쟁률을 사용한 인공신경망은 종가의 상승률이 0 이상 150퍼센트 미만인 경우의 학습이 잘 이루어져 전체 데이터에 정확도가 높은 모습을 보였으나 클래스 0과 2의 정확도가 상대적으로 낮은 결과를 확인할 수 있다. 이는 z-score 기관경쟁률을 통해 클래스 0과 2의 정확도를 높일 수 있음을 알 수 있다.

5.2 변수 중요도 분석

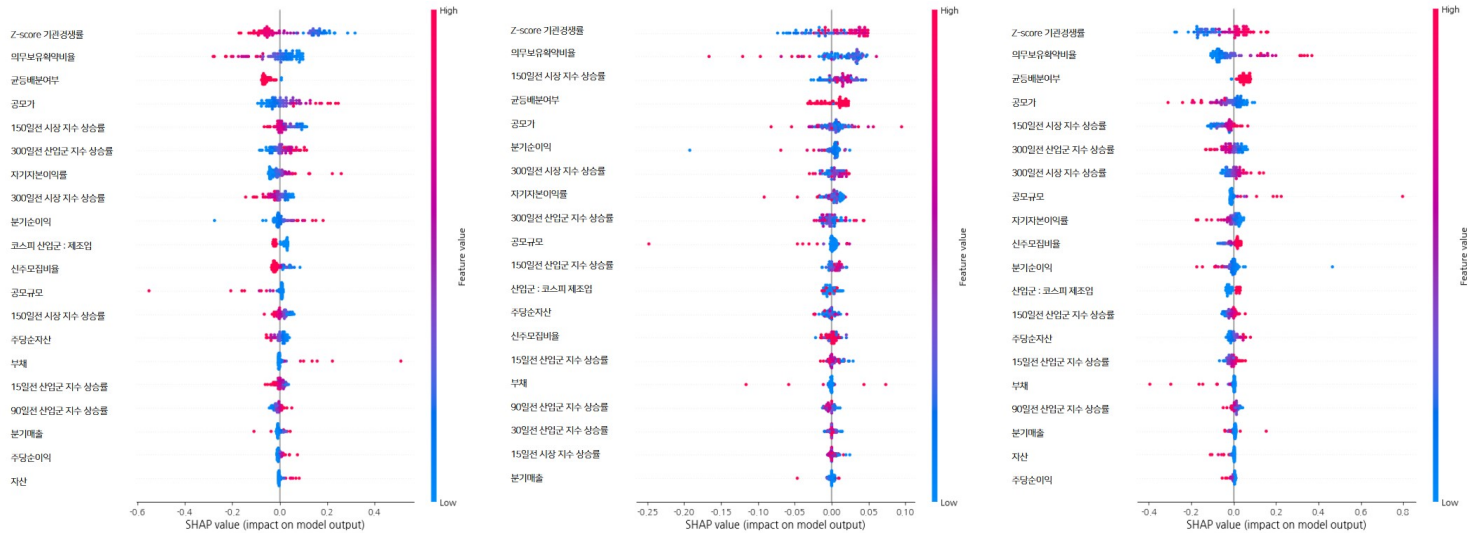
각 변수들의 SHAP 값을 그래프를 통해 시각화 하여 변수의 중요도 분석 및 예측값에 끼친 영향력 분석이 가능하다. 변수의 상대적인 중요도는 Y축을 통해 알 수 있다. Y축의 위쪽에 위치한 변수는 아래쪽에 위치한 변수들에 비해 모델이 클래스를 예측할 때 더 중요하게 작용하였다고 해석된다. X축은 Y축에 해당하는 변수의 SHAP 값으로 변수 값의 부호 및 크기에 따라 예측에 양 혹은 음의 영향을 끼쳤는지를 나타낸다. 그래프에서 각 점은 한 IPO의 Y축에 해당하는 변수를 의미하며 점의 X축에서의 위치는 해당 변수의 SHAP 값이다. 변수의 값이 클수록 빨강에 가깝고 낮은 값은 파랑에 가깝다. 각 학습 모델의 클래스별 SHAP 값 요약 그래프는 그림5.1, 5.2 그리고 5.3과 같다. 또한 각 모델 및 클래스별 변수중요도 상위 10개의 변수는 표5.6,5.7 그리고 5.8과 같다.

그림5.1, 5.2 그리고 5.3의 그래프를 통해 변수 중요도 및 변수의 값에 따라 모델의 예측에 어떻게 영향을 주었는지 해석할 수 있다. 96.3 퍼센트의 정확도를 보인 z-score 기관경쟁률을 이용하여 학습한 인공신경망이 분류한 클래스 0의 경우 z-score 기관경쟁률이 가장 중요한 변수라는 결과를 보였다. 낮은 z-score 기관경쟁률은 0보다 큰 SHAP 값을 나타냈으며 이는 기관경쟁률이 낮을수록 클래스 0으로 구분함을 의미한다. 공모가의 경우, 값이 높을수록 SHAP 값이 증가했으며 공모가가 클수록 모델이 클래스 0으로 예측하였음을 그래프5.1a에서 볼 수 있다. 기관경쟁률 둘 모두를 통해 학습한 인공신경망을 통해 예측한 정확도 72.7 퍼센트의 클래스 2 모델은 기관경쟁률이 가장 중요한 변수로 작용하였음을 그래프 5.3c를 통해 확인할 수 있다. 모델이 클래스 0을 예측하는 경우와는 반대로 기관경쟁률이 클수록, 공모가는 작을수록 양의 SHAP 값이 계산되었다. 이는 평가 데이터의 IPO가 기관경쟁률이 크고 공모가가 작을수록 클래스 2로 예측하는 경향이 있음을 의미한다. 클래스 1을 예측하는 경우 클래스 0과 2에 비하여 중요한 변수들의 SHAP 값이 크기에 따라 양수와 음수로 정확히 구분되지 않고



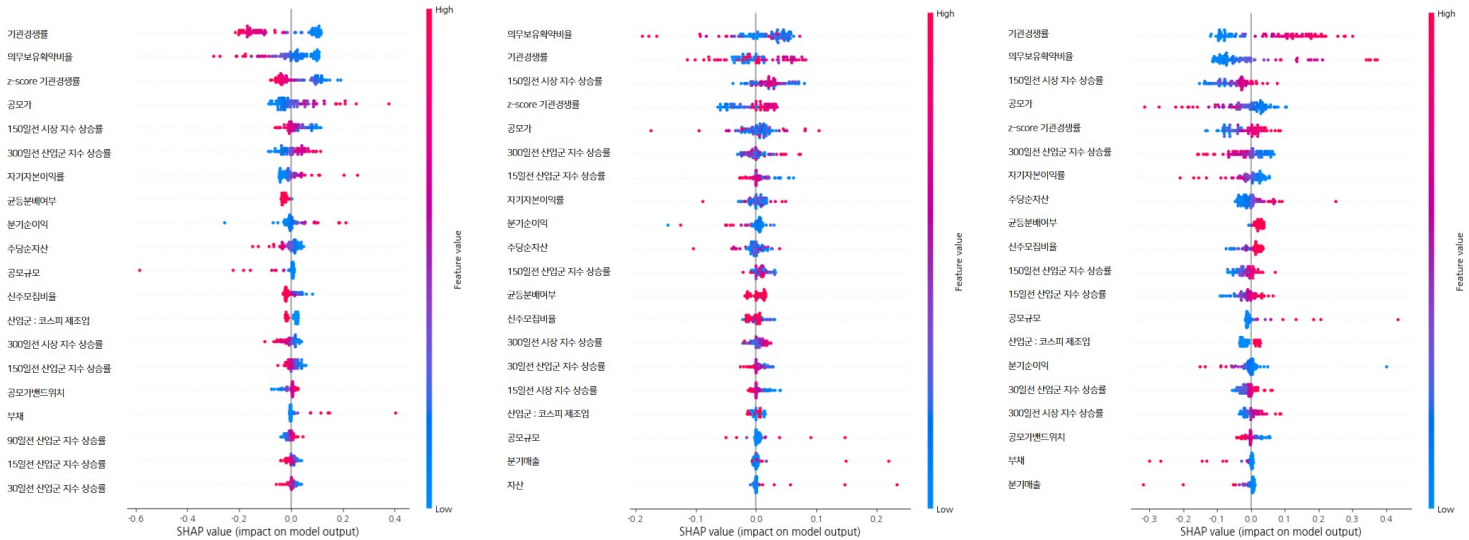
(a) Class 0 (b) Class 1 (c) Class 2

그림 5.1: 기관경쟁률 변수를 이용한 서포트 벡터 머신의 각 클래스별 SHAP 값 요약 그래프



(a) Class 0 (b) Class 1 (c) Class 2

그림 5.2: z-score 기관경쟁률을 이용한 인공신경망의 각 클래스별 SHAP 값 요약 그래프



(a) Class 0 (b) Class 1 (c) Class 2

그림 5.3: 기관경쟁률 변수 둘 모두를 이용한 인공신경망의 각 클래스별 SHAP 값 요약 그래프

섞여있는 결과를 확인할 수 있다. 또한 클래스 1과 2는 중요한 변수의 순위가 비슷한 결과를 보여주었다. 이는 학습 모델의 종가의 상승폭 예측이 종가의 등락인 클래스 0 예측에 비해 상대적으로 성능이 떨어지는 원인으로 분석된다.

그림5.1, 5.2 그리고 5.3의 그래프의 Y축에서 확인된 변수중요도를 정리한 표5.6,5.7 그리고 5.8의 결과를 통해 기관경쟁률, 의무보유확약비율 그리고 공모가와 같은 IPO 관련 정보가 각 클래스를 예측함에 있어 가장 중요한 변수로 작용함을 알 수 있다. 또한 모델이 클래스를 예측할 때 IPO 정보 외에 시장의 상황에 영향을 받음을 알 수 있다. SHAP 값을 분석한 세 모델에서 클래스 1과 2를 구분할 때 150일전 시장 지수 상승률 변수가 중요하게 작용하지만 클래스 0을 구분할 경우 상대적으로 중요도가 떨어지는 것을 확인할 수 있다. 이는 종가가 하락하는 경우 기업의 재무 상황과 IPO 정보가 시장의 상황에 비해 더 중요하게 작용함으로 분석이 가능하다. 150일의 시장 및 인덱스 상승률이 다른 기간에 비해 상대적으로 중요하게 작용한 원인으로는 그보다 짧은 기간의 경우 지표의 변화율이 적고 그보다 긴 기간의 경우 시장이 호황 및 불황인지를 보는 지표로서 너무 길기 때문인 것으로 추측된다. 균등배정 여부는 z-score 기관경쟁률을 이용한 인공신경망 모델에서 중요한 변수로 작용하였으나 학습 데이터에서 균등배정을 실행한 IPO가 적고 2021년 이후의 IPO인 평가 데이터셋에서는 모든 IPO에서 균등배정을 실행하여 중요한 변수로 작용하였다고 해석하기에는 한계가 있다. SHAP를 통해 중요한 변수가 무엇인지, 변수의 값은 모델의 IPO 결과의 예측에 어떠한 영향을 끼치는지 알 수 있으며 이는 IPO의 투자 및 자산 분배를 결정할 때 도움을 줄 수 있다.

SHAP 값 결과는 대중들이 생각하는 공모주의 상장일 상승 요인과 일치한다. 실제 IPO의 종가의 등락 및 상승폭에 영향을 주는 변수의 중요도 및 값의 크기와 SHAP 방법을 통한 분석이 일치함을 실험을 통해 확인하였다. 이는 변수중요도에서 가장 중요한 변수로 나온 공모가, 기관경쟁률 그리고 의무보유확약을 통해 확인이 가능하다.

		RF	XGB	LR	SVM	ANN
기관경쟁률	0	80%	87%	83%	87%	89%
	1	66%	59%	25%	14%	47%
	2	75%	67%	83%	92%	45%

표 5.4: 기관경쟁률, 의무보유확약비율, 공모가만을 이용한 학습의 각 클래스별 정확도

	RF	XGB	LR	SVM	ANN
정확도	71.3%	68.3%	49.5%	44.6%	59.0%

표 5.5: 기관경쟁률, 의무보유확약비율, 공모가만을 이용한 학습의 정확도

각 변수들의 SHAP 값은 기존 IPO의 흥행을 분석하는 연구들에서 중요하게 다루었으며, 변수들의 값에 따른 IPO의 흥행 여부는 SHAP 그래프를 통해 가능한 변수의 값이 예측값에 영향을 주는 방향과 일치한다. IPO의 흥행에 영향을 미치는 공모가의 중요성과 값은 저평가와 관련이 있다. IPO의 흥행을 결정하는 저평가는 IPO를 하고자 하는 기업 및 주관사가 가격을 잘못 측정하여 높은 가격에 주식을 상장하는 경우 달성되지 않는다. 이는 단기간으로는 흥행의 실패를 야기하며 상장일 종가가 공모가에 비에 하락하는 원인이 된다.[3] 높은 기관경쟁률은 기관 및 일반 청약자의 정보의 비대칭성과 투기 거품으로 인해 단기적으로 IPO의 흥행에 긍정적인 영향을 끼친다. [1] 즉, 높은 기관경쟁률은 상장일 종가가 공모가 대비 상승하는 요인이다. 의무보유확약비율 또한 높을수록 IPO의 종가의 상승 요인이 된다.[2] 의무보유확약비율은 기관투자자들이 락업(Lock-up) 조건을 통해 더 많은 주식을 청약하는 경우 상승한다. 이는 상장일 주식의 유통량을 제한하여 투자자의 과잉된 확신이 공모시장을 쉽게 지배할 수 있도록 하며, 최종적으로 종가의 상승을 불러온다.[29] SHAP 값 실험을 통해 변수중요도가 높은 기관경쟁률,공모가 의무보유확약비율은 공모가 대비 종가의 등락을 결정하는 주요한 원인이며, 기관경쟁률과 의무보유확약비율이 높고 공모가가 낮을수록 공모가 대비 종가가 상승할 가능성이 크며 이는 IPO 사례 및 사전 연구 결과와 일치한다.

앞 세 모델의 변수중요도 표를 통해 중요도가 높고 IPO 관련 정보에 해당되는 기

관경쟁률, 의무보유확약비율 그리고 공모가 총 3개의 변수를 사용하여 랜덤 포레스트, XGBoost, 로지스틱 회귀, 서포트 벡터 머신 그리고 인공신경망에 대한 추가 실험을 진행하였다. 실험 결과는 표5.4, 표5.5와 같다. 세 개의 변수만을 통해 학습한 경우 모든 모델에서 80퍼센트 이상의 클래스 0 예측 정확도를 보였다. 랜덤 포레스트 모델에서 71.3퍼센트의 가장 좋은 정확도를 보였으며 각 클래스별 정확도 또한 60퍼센트 이상의 정확도 결과를 보여주었다. 랜덤 포레스트 모델의 경우 적은 수의 변수를 사용하였을 때 학습 데이터에 대한 과적합이 적어 모든 변수를 사용한 학습 대비 성능이 향상된 결과를 보여주었다. 세 개의 변수만을 수집하여 인공신경망과 랜덤포레스트를 학습하면 보다 적은 수의 변수와 짧은 학습 시간으로 정확도 80퍼센트 이상의 상장일 공모가 대비 종가의 등락 예측 성능을 보인다.

	Class 0	Class 1	Class 2
1	기관경쟁률	기관경쟁률	의무보유확약비율
2	의무보유확약비율	150일전 시장 지수 상승률	150일전 시장 지수 상승률
3	공모가	의무보유확약비율	기관경쟁률
4	분기순이익	자기자본이익률	300일전 산업군 지수 상승률
5	공모규모	30일전 시장 지수 상승률	공모가
6	300일전 산업군 지수 상승률	300일전 산업군 지수 상승률	30일전 시장 지수 상승률
7	30일전 산업군 지수 상승률	주당순이익	30일전 산업군 지수 상승률
8	90일전 산업군 지수 상승률	공모가	균등배정여부
9	공모가밴드위치	균등배정여부	60일전 산업군 지수 상승률
10	90일전 산업군 지수 상승률	15일전 시장 지수 상승률	90일전 산업군 지수 상승률

표 5.6: 기관경쟁률 변수를 이용한 서포트 벡터 머신의 각 클래스별 중요변수 순위

	Class 0	Class 1	Class 2
1	z-score 기관경쟁률	z-score 기관경쟁률	z-score 기관경쟁률
2	의무보유확약비율	의무보유확약비율	의무보유확약비율
3	균등배정여부	150일전 시장 지수 상승률	균등배정여부
4	공모가	균등배정여부	공모가
5	150일전 시장 지수 상승률	공모가	150일전 시장 지수 상승률
6	300일전 산업군 지수 상승률	분기순이익	300일전 산업군 지수 상승률
7	자기자본이익률	300일전 시장 지수 상승률	300일전 시장 지수 상승률
8	300일전 시장 지수 상승률	자기자본이익률	공모규모
9	분기순이익	300일전 산업군 지수 상승률	자기자본이익률
10	산업군 : 코스피 제조업	공모규모	신주모집비율

표 5.7: z-score 기관경쟁률 변수를 이용한 인공신경망의 각 클래스별 중요변수 순위

	Class 0	Class 1	Class 2
1	기관경쟁률	의무보유확약비율	기관경쟁률
2	의무보유확약비율	기관경쟁률	의무보유확약비율
3	z-socre 기관경쟁률	150일전 시장 지수 상승률	150일전 시장 지수 상승률
4	공모가	z-score 기관경쟁률	공모가
5	150일전 시장 지수 상승률	공모가	z-score 기관경쟁률
6	300일전 산업군 지수 상승률	300일전 산업군 지수 상승률	300일전 산업군 지수 상승률
7	자기자본이익률	15일전 산업군 지수 상승률	자기자본이익률
8	분기순이익	자기자본이익률	주당순이익
9	주당순자산	분기순이익	균등배정여부
10	공모규모	주당순이익	신주모집비율

표 5.8: 기관경쟁률 변수 둘 모두를 이용한 인공신경망의 각 클래스별 중요변수 순위

제 6 장 결론

본 논문에서는 기업 정보, 시장상황 IPO관련 정보들과 생성한 파생변수들을 이용하여 한국주식시장 IPO의 상장일 주가를 기계 학습 모델을 통해 예측하였다. IPO의 주가를 예측함으로써 개인투자자들이 종가가 하락하는 청약을 회피하고 상승율이 더 높은 IPO에 많은 청약증거금을 할당할 수 있도록 도움을 주는 모델을 만들고자 하였다. 실험을 통해 전체 데이터에 대한 정확도 71.6 퍼센트, 클래스 0 정확도 92.6 퍼센트 그리고 클래스 2 정확도 72.7 퍼센트의 성능을 보인 인공신경망 모델을 제시하였다. 또한 SHAP 기법을 통해 기준 성능을 만족한 세 모델들에 대한 변수들의 SHAP값을 계산하였다. 계산된 SHAP 값을 그래프로 시각화하여 변수중요도와 변수의 값이 모델의 예측에 미치는 영향을 분석하였다. SHAP를 통해 기존 연구들에는 없던 변수의 중요도 분석과 값에 따른 공모주의 등락에 대한 해석이 가능하였으며, IPO의 단기적인 흥행 및 주가 예측에 있어 과거 사례 및 사람들의 연구와 일치하는지를 확인하였다. 중요한 변수가 무엇인지, 변수의 값이 IPO의 흥행에 어떠한 영향을 끼치는 지를 알 수 있으며 이는 IPO의 투자의 참여 및 자산 분배에 있어 정보를 제공한다. 마지막으로 가장 중요한 변수 3개를 통한 학습을 실험하였고 종가의 등락 예측에 있어 80퍼센트 이상의 정확도를 보임을 알 수 있었다.

IPO의 상장일 기준 가격제한폭이 ±30 퍼센트로 변경된 2015년 6월 15일부터 2022년 7월까지 유가증권시장과 코스피 시장에 상장된 IPO 데이터를 이용하여 실험하였다. 시간이 지남에 따라 IPO 시장이 과열되어 절대적인 기관경쟁률 수치는 IPO의 등락 및 상승률을 정확하지 반영하지 못할 수 있기 때문에 이를 보완하고자 z-score 기관경쟁률을 생성하여 학습을 진행하였다. 기관경쟁률 및 z-score 기관경쟁률이 모델의 학습의

영향을 끼치는지를 분석하기 위하여 각 변수를 따로 사용하거나 둘 모두를 사용하는 3개의 경우로 나누어 학습을 진행하였다. 랜덤 포레스트 그리고 XGBoost와 같은 의사결정나무 기반 앙상블 기법과 로지스틱 회귀, 서포트 벡터머신 그리고 인공신경망 모델을 통해 학습을 진행하였다. 학습 후 기준 이상의 성능을 가진 학습 모델에 한정하여 각 알고리즘에 최적화된 SHAP 기법을 통해 변수가 모델의 예측에 미치는 영향 및 중요도에 대하여 분석하였다.

전체 데이터에 대하여 정확도가 70퍼센트 이상, 각 클래스에 대하여 정확도가 60퍼센트 이상인 모델은 기관경쟁률을 이용한 서포트 벡터 머신, z-score 기관경쟁률을 이용한 인공신경망 그리고 둘 모두를 이용한 인공신경망이다. 서포트 벡터 머신 모델은 클래스 0 예측의 정확도가 80퍼센트로 92퍼센트 이상의 정확도 결과를 보여준 인공신경망 모델에 비해 공모가 대비 종가의 등락을 잘 예측하지 못하였다. 두 인공신경망 모델은 공모가 대비 종가의 상승률 예측 정확도는 약 61퍼센트로 종가가 하락을 예측함보다 성능이 낮았다. z-score 기관경쟁률을 통해 학습한 인공신경망은 기관경쟁률만을 사용하여 학습한 인공신경망보다 종가의 등락인 클래스 0과 종가가 크게 상승한 클래스 2의 예측에 있어 더 높은 정확도를 보여주었다. 전체 데이터에 대한 정확도 71.6 퍼센트, 클래스 0 정확도 92.6 퍼센트 그리고 클래스 2 정확도 72.7 퍼센트의 성능을 보인 두 기관경쟁률 모두를 이용한 인공신경망 모델을 최종 예측 모델로 선택하였다.

성능 기준을 만족한 세 통계학습모델에 대하여 서포트 벡터 머신은 커널 SHAP 방식을, 인공신경망은 딥 SHAP 방식을 통해 SHAP 값을 계산하였다. SHAP 값의 계산 및 시각화를 통해 각 모델에서 어떠한 변수가 중요하고 변수의 값에 따라 모델이 IPO의 등락 및 상승율을 예측함에 있어 어떠한 영향을 받았는지를 확인하였다. 세 모델 모두에서 기관경쟁률, 의무보유확약비율 그리고 공모가가 중요한 변수로 작용하였다. 기관경쟁률 둘 모두를 사용한 인공신경망의 경우 z-score 기관경쟁률보다 기관경쟁률이

모든 클래스에서 더 중요하게 작용하였다. 클래스 0의 경우 모델이 예측할 때 다른 클래스에 비하여 시장 지수 상승률의 중요도가 낮은 모습을 볼 수 있고 이는 공모가 대비 종가의 하락에는 IPO 정보가 시장의 호황 및 불황보다 더 중요하게 작용함을 알 수 있다. 클래스 1과 2는 기관경쟁률, 의무보유확약비율, 공모가 및 150일 시장 지수 상승률이 중요한 변수들로 작용하였고 중요도 순위 또한 큰 차이가 없었다.낮은 기관경쟁률, 낮은 의무보유확약비율 그리고 높은 공모가를 가진 IPO를 클래스 0으로 구분하고, 반대로 높은 기관경쟁률, 높은 의무보유확약비율 그리고 낮은 공모가를 가진 IPO를 모델이 클래스 2로 예측함을 SHAP 값의 그래프를 통해 확인이 가능하였다. 클래스 1의 경우 클래스 SHAP 값의 분포가 클래스 2의 분포와 비슷하나 하나의 SHAP 값을 경계로 값에 따라 데이터가 정확히 구분되지 않는 모습을 그래프에서 보였으며 이는 모델이 종가의 상승폭 예측이 등락의 예측에 비해 성능이 떨어지는 원인으로 생각된다. 기관경쟁률, 의무보유확약비율 그리고 공모가가 주요한 IPO 흥행의 성패를 가르는 요인으로 나온 SHAP 방법을 통한 변수 중요도 순위 및 변수의 값에 따른 해석은 과거 IPO의 사례에서 보여주었던 IPO 흥행의 실패의 요인과 일치한다. 즉, 통계학습기법을 이용한 한국주식시장 IPO의 주가 예측 모델에서 SHAP 방법을 통한 변수의 설명이 IPO 사례 및 연구 결과와 일치함을 알 수 있다. 이는 기존 연구들과 같이 기계학습을 통해 단순히 예측하는 것이 아닌 어떠한 변수가 IPO의 흥행에 주요한 영향을 끼치는 지 알 수 있다. SHAP을 통해 알 수 있는 변수의 중요도와 그 값에 따른 영향에 대한 정보는 IPO의 참여 여부 및 자산 분배 정도를 결정하는 것에 도움을 줄 수 있다. 중요도 순위가 높은 공모가, 기관경쟁률 그리고 의무보유확약비율 3개의 변수만을 이용하여 실험을 진행하였으며 중요도가 높은 변수들을 통해 종가의 등락 예측 정확도나 높으나 주가의 상승폭은 상대적으로 잘 구분하지 못하는 결과를 확인할 수 있었다.

본 논문의 한계점으로는 먼저 21년 이후 데이터의 수가 적어 IPO의 균등분배여부

가 IPO의 흥행에 끼친 영향을 해석함에 제한적인 측면이 있다는 것이다. 학습 및 평가 데이터가 임의가 아닌 시기로 나누어져 학습 데이터에 포함된 균등배정 제도가 적용된 IPO의 수는 매우 적었으며 평가 데이터에서는 모든 IPO에서 균등배정 제도가 적용되어 변수가 예측값에 끼치는 영향의 분석에 왜곡이 발생할 수 있는 측면이 있다. 또한 IPO 데이터의 수가 7년간 약 500개로 그 수가 매우 적어 데이터를 학습, 검정 그리고 평가로 나누는 것에 한계가 있었고 이는 통계학습모델이 과적합 되었을 가능성이 있다. 하지만, SHAP 방법을 통해 다양한 변수들이 IPO의 흥행에 어떠한 영향을 끼치는지 살펴보았고 실제 사람들의 생각과 비슷하다는 것을 확인할 수 있다는 점에 의의를 둘 수 있다.

본 논문의 후속 연구 과제로는 한국 주식시장이 아닌 주식의 상하한이 없는 미국 주식시장에서 상장 기업의 적절한 주가를 찾는 공모 적정가 회귀분석 연구 및 본 논문에서 사용한 정보 외 언론 및 소셜 네트워크 서비스(SNS)의 언어정보를 이용하여 IPO의 흥행 여부를 예측하는 모델 생성이 있다.

참고 문헌

[1] S. Agarwal, C. Liu, and S. G. Rhee, *Investor demand for ipos and aftermarket performance: Evidence from the hong kong stock market*, Journal of International Financial Markets, Institutions and Money, 18 (2008), pp. 176–190.

[2] J. D. Arthurs, L. W. Busenitz, R. E. Hoskisson, and R. A. Johnson, *Signaling and initial public offerings: The use and impact of the lockup period*, Journal of Business Venturing, 24 (2009), pp. 360–372.

[3] C. B. Barry and R. H. Jennings, *The opening price performance of initial public offerings of common stock*, Financial Management, (1993), pp. 54–63.

[4] E. Basti, C. Kuzey, and D. Delen, *Analyzing initial public offerings' short-term performance using decision trees and svms*, Decision Support Systems, 73 (2015), pp. 15–27.

[5] S. Basu, *Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis*, The journal of Finance, 32 (1977), pp. 663–682.

[6] R. P. Beatty and J. R. Ritter, *Investment banking, reputation, and the underpricing of initial public offerings*, Journal of financial economics, 15 (1986), pp. 213–232.

[7] L. Breiman, *Random forests*, Machine learning, 45 (2001), pp. 5–32.

[8] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, et al., *Xgboost: extreme gradient boosting*, R package version 0.4-2, 1 (2015), pp. 1–4.

[9] E. Chong, C. Han, and F. C. Park, *Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies*, Expert Systems with Applications, 83 (2017), pp. 187–205.

[10] C. Cortes and V. Vapnik, *Support-vector networks*, Machine learning, 20 (1995), pp. 273–297.

[11] E. F. Fama and K. R. French, *Size and book-to-market factors in earnings and returns*, The journal of finance, 50 (1995), pp. 131–155.

[12] J. H. Friedman, *Greedy function approximation: a gradient boosting machine*, Annals of statistics, (2001), pp. 1189–1232.

[13] D. Gale and L. S. Shapley, *College admissions and the stability of marriage*, The American Mathematical Monthly, 69 (1962), pp. 9–15.

[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*, MIT press, 2016.

[15] J. J. Han and H.-j. Kim, *Stock price prediction using multiple valuation methods based on artificial neural networks for kosdaq ipo companies*, Investment Analysts Journal, 50 (2021), pp. 17–31.

[16] H. He and E. A. Garcia, *Learning from imbalanced data*, IEEE Transactions on knowledge and data engineering, 21 (2009), pp. 1263–1284.

[17] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*, vol. 398, John Wiley & Sons, 2013.

[18] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, arXiv preprint arXiv:1412.6980, (2014).

[19] S. Lipovetsky and M. Conklin, *Analysis of regression in game theory approach*, Applied Stochastic Models in Business and Industry, 17 (2001), pp. 319–330.

[20] S. M. Lundberg, G. G. Erion, and S.-I. Lee, *Consistent individualized feature attribution for tree ensembles*, arXiv preprint arXiv:1802.03888, (2018).

[21] S. M. Lundberg and S.-I. Lee, *A unified approach to interpreting model predictions*, Advances in neural information processing systems, 30 (2017).

[22] D. C. Montgomery and G. C. Runger, *Applied statistics and probability for engineers*, John wiley & sons, 2010.

[23] J. Patel, S. Shah, P. Thakkar, and K. Kotecha, *Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques*, Expert systems with applications, 42 (2015), pp. 259–268.

[24] D. Quintana, Y. Sáez, and P. Isasi, *Random forest prediction of ipo underpricing*, Applied Sciences, 7 (2017), p. 636.

[25] A. Shrikumar, P. Greenside, and A. Kundaje, *Learning important features through propagating activation differences*, in International conference on machine learning, PMLR, 2017, pp. 3145–3153.

[26] 강민석, 윤용건, and 권대중, *부동산투자회사 (reits) 의 투자 활성화 방안에 관한 연구 (일반 투자자의 투자의식을 중심으로)*, 대한부동산학회지, 29 (2011), pp. 77–106.

[27] 곽준희, *코로나 19 전후 개인투자자의 손익률 변화와 시사점*, 주간금융브리프, 31 (2022), pp. 3–8.

[28] 김범준, *자본시장법상 기업인수목적회사 (spac) 제도의 도입과 법제적 과제*, 기업법연구, 24 (2010), pp. 119–148.

[29] 변진호 and 조성순, *Ipo 락업 제도가 저가발행현상에 미치는 영향*, 한국증권학회지, 40 (2011), pp. 405–429.

[30] 신상훈, 이현준, and 안재준, *비재무 정보를 활용한 ipo 주식의 상장일 가격 등락 예측에 관한 연구*, 한국데이터정보과학회지, 29 (2018), pp. 425–439.

[31] 조득환, 류호선, 정승환, and 오경주, *인공지능 (ai) 을 활용한 공모주 투자여부 및기준 수익률 달성 여부 예측 모델*, 한국데이터정보과학회지, 31 (2020), pp. 579–590.

Abstract

Listing Day Price Prediction of Korean Stock Market IPO by Machine Learning

Jongu Lee
Department of Industrial Engineering
The Graduate School
Seoul National University

In Korean Exchange, Initial Public Offering(IPO) means a non-listed company listing its stock on the stock exchange. Investing in offering shares is a popular investment method after 2020, most of IPOs show rising closing price due to underpricing. However, some of the IPOs' closing price might fall and result in losses and if the schedule for the IPOs overlaps, investors have to determine the distribution of assets. This study proposes a machine learning model that predicts listing-day closing price trained with finance, market situation, IPO information variables, and derived variables like z-score oversubscription. In addition, by using SHAP(Shapley Additive Explanation), a way of Explainable Artificial Intelligence, this paper analyzes the importance and influence of the value of variables. The artificial neural network(ANN) model trained by Korean Exchange market IPO data and derived variables showed 92.6 percent accuracy of up and down and 72.7 percent accuracy of closing price rising more than 150 percent. By SHAP, we can know that significant variables are offering price, oversubscription, and lock-up rate. Smaller offering prices, lower over-

subscription, and higher lock-up rate make the model predict that the closing price is less likely to fall and will rise higher, which is equal to past examples and previous studies.

Keywords: IPO, Machine Learning, Explainable Artificial Intelligence, SHAP

Student Number: 2021-29800