

**공모가 시초가 예측**

* **[경영 빅데이터 분석] 5조 팀 프로젝트 -**



|  |  |  |
| --- | --- | --- |
| 팀원 | 학번 | 전공 |
| 고도환 | 5434535 | 경영정보학전공 |
| 전시영 | 4601263 | 경영정보학전공 |
| 정한별 | 5533017 | 컴퓨터공학전공 |
| 풍티하이하 | 5731440 | 한국어교육전공 |
| 김우준 | 5605627 | 경영정보학전공 |

**공모주 시초가 예측**

**-코스닥시장에서의 공모주 시초가 예측-**

**-요 약-**

코로나19 이후 IPO시장은 상장 첫날 수익률이 2020년과 2021년 평균 53.8%와 48.1%로 높은 수치를 보여주고 있다. 이에 공모주에 대해 개인의 투자가 활발히 이루어지고 있지만, 개인투자자들이 공모주 시장에서 얻을 수 있는 정보는 매우 한정되어 있다. 시중에 공개된 데이터로부터 인공지능과 통계학적 방법론을 통하여 공모주 시초가가 더블이 되는 것을 예측하는데 유의미한 변수를 찾고, 이로 인해 투자자의 의사결정에 도움을 줄 수 있는 결과를 얻을 수 있었다.

**-서론 및 관련 연구-**

***“석 달 평균 수익률 139%, 지난해 평균 수익률 38.6%”***   
인터넷에 ‘공모주 수익률’이라고 검색했을 때 나오는 기사 제목들 중 일부이다.   
이처럼 우리나라에서 공모주 투자 수익률이 높다는 이야기가 여기저기서 들리기 시작하면서 많은 사람이 공모주 투자에 관심을 기울이게 되었다. 과거에는 기관 투자자에 더하여 소수의 투자자들이 큰 규모의 자본과 많은 인력을 투입하여 일반 개인 투자자들에 비해 유리한 환경 속에서 투자전략을 수립하여 공모에 참여하였지만, 현재는 정보통신의 발달로 인해 투자정보제공의 접근성이 좋아짐에 따라 일반 투자자들도 공모주 투자에 참여할 수 있는 기회가 점점 더 증가하는 추세이다.

이번 보고서는 ‘선행연구 투자주체별 청약경쟁률과 공모주(IPO)수익률에 관한 연구(김종근, 2016)’, ‘공모주 청약경쟁률과 IPO 이후 주가행태에 관한 연구(백삼열, 2019)’, ‘인공지능(AI)을 활용한 공모주 투자여부 및 기준 수익률 달성 여부 예측 모델(조득환 외 3인, 2020)’ 그리고 ‘머신러닝 알고리즘을 활용한 신규공모주 수익률 예측(최연지,2022)’을 참고하여 검증된 변수들과 가설을 통한 추가 변수들을 선정하여 신규 상장된 종목이 첫 거래일 공모가 대비 두 배로 시초가가 형성될지, 안 될 지를 ‘이진 분류’로 시행해 보았다.

보고서는 다음과 같이 구성된다.

**[목차]**  
**1. 분석 준비** - 공모주 시장의 의미와 데이터 수집 과정

**2. 데이터 전처리**- 종속변수 선정 및 데이터[missing value] 처리, 추가로 필요한 독립 변수 생성

**3. 가설 설정 및 검증**

- 전체 변수를 활용하여 가설을 세우고 가설 검증 진행

**4. 변수의 시사점 도출**

- 검증된 가설의 유의미한 변수를 활용한 CART-LR, C5.0-LR 모델 생성

- 가장 많은 영향을 미치는 변수의 유의확률을 가지는 지점 확인 및 시사점 분석

**5. 변수의 유의미한 지점 도출**

- 로지스틱 회귀곡선의 회귀선을 통한 6개 변수의 유의미한 지점 확인

**6. 최종모델 구현**

- 인공신경망을 통한15개의 변수로 최종 모델을 학습을 진행하여 최종모델 구현 **7. 최종 결론과 한계점**

**1. 분석 준비**

**1.1. 공모주 시장**

공모주 시장은 새로운 기업이 주식을 처음으로 공개하고 증권거래소에 상장하는 과정을 의미한다.   
기업들이 기업공개(IPO)를 통해 증권거래소에 상장하게 된다. 이는 기업이 추가 자금을 조달하고, 투자자들에게는 주식을 제공하는 기회를 제공한다.

공모주 시장에서 기업은 주식 발행을 위해 관련된 법률 절차와 규정을 준수해야 한다. 기업은 주식 발행을 위한 상장을 신청하고, 상장 전 시장 감시 기관의 검토와 승인을 받아야 한다. 발행 계획이 확정되면, 기업은 주식 발행을 위한 공고를 신문 등에 게재하여 공시한다. 이로써 투자자들에게 주식 공모에 대한 정보를 알리게 되며, 공고 이후, 주식 발행을 위한 IPO가 도래하게 된다. 이 기간 동안에 투자자들은 공모 주식을 구매할 수 있으며, 구매 방식은 일반 공모와 기관 투자자 공모 등이 있다. 공모 기간이 종료가 되면, 기업은 투자자들로부터 받은 주문을 검토하여 주식을 확정하고, 주식의 배정을 진행한다.

텍스트, 폰트, 스크린샷, 로고이(가) 표시된 사진

자동 생성된 설명

그림 1. IPO 과정

**1.2. 데이터 수집**

데이터는 공모주 대표사이트인 ‘38커뮤니케이션’과 ‘ipostock’에서 [BeautifulSoup]를 사용하여 크롤링하여 가져왔다. 데이터의 범위는 금융당국이 전일 종가 대비 가격제한폭을 ±30%로 확대한 2015년 6월 15일부터 현재인 2023년 5월까지의 8년치 코스닥 공모주 데이터 528개를 가져왔으며, 가져온 칼럼은 다음과 같다.

|  |  |  |  |
| --- | --- | --- | --- |
| **신규상장일** | **기업명** | **공모가** | **시초가** |
| 첫날종가 | 현재가 | 희망공모가\_하한 | 희망공모가\_상한 |
| 확정공모가 | 기관경쟁률 | 청약경쟁률 | 의무보유확약 |
| 주간사 | 신주모집 | 구주매출 | 시초/공모(%) |
| 시장구분 | 업종 | 기업구분 | 매출액 |
| 순이익 | 자본금 | 우리사주조합 | 기관투자자등 |
| 일반청약자 | 공모후상장주식수 (주식수) | 유통가능주식합계 (주식수) | 유통가능주식합계 (비율) |

표 1. 크롤링 칼럼

**2. 데이터 전처리**

**2.1. 종속변수 선정**

과거 공모주 시장에서의 데이터를 바탕으로 종속변수와 독립변수를 선택하였다. 종속 변수는 IPO 첫날 시초가와 공모가의 손익을 기준으로 설정하는데, 본 보고서에서는 공모가 대비 시초가 수익률 ‘100%’를 기준으로 삼기로 하였다. 한 가지의 대표적인 이유를 들자면, 한 증권회사의 공모주 청약란을 보면 소위 말하는 ‘따’ 와 ‘따상’을 달성할 수 있는 지에 대한 확률을 제공한다. ‘따’는 공모가 대비 시초가의 수익률이 ‘100%’이고 ‘따상’은 두 배로 시초가가 형성된 뒤에 가격 제한 폭인 +30%까지 상승해 마감하는 것을 뜻한다. 그만큼 ‘100%’ 라는 수치가 중요하다고 생각이 되어 설정하게 되었다.

**2.2. missing value 처리**

‘38커뮤니케이션’과 ‘ipostock’에서 데이터를 가져와서 ‘기업명’을 기준으로 [Join]하여 하나의 csv파일로 만들 필요가 있었고, 그 과정에서 누락된 값이 발생하였는데 발생 이유는 다음과 같다.

|  |  |  |
| --- | --- | --- |
| **38커뮤니케이션** | **ipostock** | **비고** |
| 대모(구.대모엔지니어링) | 대모엔지니어링 | 불일치 |
| 아주IB투자 | 아주아이비투자 | 한/영 |
| 제이티씨(JTC) | JTC | 한/영 |
| 하나머티리얼즈 | 하나실리콘 | 상호변경 ipostock에서 데이터 없음 |
| 디에스씨인베스트먼트 | DSC인베스트먼트 | 한/영 |
| 코썬바이오 | 현성바이탈 | 상호변경 |

표 2. missing value

몇몇 기업은 매출액, 순이익, 자본금 등이 표기가 안되어 공백으로 표시된 것이 있었고, 외국계 기업의 달러나 엔화, 위안화 같은 경우 38커뮤니케이션에서 정보를 아예 기재를 안 하는 경우도 있었다. 그래서 ‘ipostock’라는 사이트에서 추가로 정보를 찾아서 교차검증을 진행하였다.

**2.3. 변수 추가**

비상장 기업이 상장을 하는 날의 시초가가 결정되기까지 많은 변수들과 요인들이 존재한다. 그래서 이러한 요인들을 데이터로 사용할 수 있는 정보로 만들기 위해 수치화 하였다. 상장시기는 공모주식의 상장월을 의미하며, 최근 공모주 동향1과 2는 각각 상장일을 기준으로 3개월 내의 다른 공모주식에 대하여 그 공모주식이 시초가/공모가의 값이 ‘100%’, ‘양수’가 몇 개나 있는지를 나타낸다. 최근 "따" 비율과 최근 “양수” 비율은 각각 공모주 동향 1, 2를 최근 3개월 공모주 수로 나눈 값이고, 업종 같은 경우는 한국표준산업분류(10차)기준으로 대분류에 해당하는 값으로 재분류를 해주었고, IPO 주간 순위는 “더 벨” IPO 주관 순위 바탕으로 상장일 기준 전년도 주관 순위 입력하여 주간사를 대체하였다. 밴드수익률은 결정된 공모에서 주관사의 희망 공모가격의 상단을 나눈 비율이고 1보다 크다면 밴드 상단 초과이다. 구주매출의 경우 구추매출에 대한 비율을 구해 구주매출(비율)로 대체하였다.

**3. 가설 설정 및 검증**

**3.1. 가설 설정**

가설은 총 28개의 변수를 통해서 설정할 것이다. 변수는 [경영 빅데이터 분석] 수업에서 사용한 변수 15개를 기본으로 전제로 하였고, 추가로 논문이나 각종 자료를 통해 유의미한 변수들을 찾는 작업을 할 것이다.   
기존 15개의 변수는 다음과 같다.   
상장 시기(상장월), 공모가, 기관경쟁률, 청약경쟁률, 의무보유확약, 대표주간사, IPO 주간사 순위, 총 공모주식 수, 구주매출(비율), 매출액(백만원), 순이익(백만원), 자본금(백만원), 우리사주조합(비율), 공모주 상장주식수(주식수), 유통가능주식합계(비율)

여기에 우리는 다음과 같은 가설을 추가하였다.   
가설1: 최근 주식시장 동향이 공모주 시초가 예측에 영향을 미칠 것이다.   
가설2: 밴드수익률(즉, 공모가격이 희망 공모가 상한에 얼마나 가까운지)이 공모주 시초가 예측에   
 영향을 미칠 것이다.   
가설3: 전문투자자 비율이 공모주 시초가 예측에 영향을 미칠 것이다.   
가설4: 자기자본 순이익률이 공모주 시초가 예측에 영향을 미칠 것이다.

이렇게 총 4가지의 가설로 추가 변수를 4개를 설정하였다.

정리하자면, 기본 변수 15개와 추가 변수 4개로 총 19개의 변수를 통해서 의사결정나무의 CART, C5.0로 학습하여 추가변수 중 유의미한 변수가 무엇인지 확인하여 가설이 맞는지에 대해 검증을 진행할 것이다.   
그러고 나서 이를 바탕으로 로지스틱 회귀 모형을 학습하여 추가한 변수 및 기존 변수가 공모주 시초가 100% 달성에 얼마나 영향을 미치는지를 확인하고자 한다.

아래의 표는 추가4가지 변수를 어떤 이유로 선정하게 되었는지 좀 더 자세하게 볼 수 있다.

|  |  |
| --- | --- |
| **추가 변수** | **설명** |
| 최근 “따” 비율 | 최근 “따” 비율이란 최근 3개월 공모주 수 대비 공모주 시초가가 더블이 된 변수를 의미 → 최근 주식시장의 동향을 학습해 공모주 시초가 100% 달성에 유의미하게 작용할 것으로 예상된다.  (※ 3개월을 기준으로 잡은 이유는 주관사가 적정 공모가를 결정하고 희망 공모가를 선정하는데 일반적으로 3~4개월 걸리기 때문) |
| 밴드 수익률 | 공모주 시초가 예측은 “공모가/희망공모가\_상한”으로 공모가격이 희망 공모가 상한에 얼마나 가까운지 측정 (1에 가까울수록 희망 공모가 상한에 가깝다는 것을 의미) → 이러한 변수는 투자자들의 기대심리를 반영. 즉, 투자자들이 그 주식에 대해 높은 가격을 지불할 준비가 되었음을 나타내기에 공모주 시초가 100%달성에 유의미하다고 예상된다. |
| 전문 투자자 비율 | 전문 투자자는 정보에 접근성에 일반 투자자와 차이가 있다. 전문 투자자는 더 많은 경험과 전문지식을 갖추고 있기에, 전문 투자자의 비율이 높으면 일반 투자자들에게 기대심리를 주어 더욱 활발한 투자를 불러일으킬 수 있다. 그래서 공모주 시초가 100% 달성에 유의미할 것으로 예상된다. |
| 자기자본 순이익률(ROE: Return On Equity) | ROE는 자본 대비 얼마나 많은 이익을 얻었는지 나타낸다. 이는 회사의 경영 능력과 재무 건전성이 우수함을 보여줄 수 있기 때문에 ROE가 높으면 투자자들이 안정감을 느끼기에 공모주 시초가 100%를 달성하는데 유의미한 변수가 될 것으로 예상된다. |

표 3. feature selection

**3.2. 가설 검증**

가설 검증은 의사결정나무의 CART, C5.0을 통해서 진행하였다. 이 결과 CART를 통해서는 추가 변수 중 최근 “따” 비율, 유의미하다는 것을 확인할 수 있었다. C5.0을 통해서는 위의 최근 “따” 비율, 전문 투자자비율에 더하여 밴드수익률까지 유의미하다는 것을 확인할 수 있었다. 결론적으로 가설 설정에 사용된 변수 중 ROE 빼고는 모두 공모주 시초가 예측에 영향을 끼치는 유의미한 변수이다.

**4. 변수의 시사점 도출**

**4.1. 개요**

CART-LR, C5.0-LR 모델을 통해서 증명된 가설의 추가 변수에 대한 시사점을 분석했다.   
CARR-LR은 기본 변수 15개 + 최근 “따” 비율, 전문투자자비율을 추가하여 학습하였고, C5.0-LR은 기본 변수 15개에 + 최근 “따” 비울, 전문투자자비율, 밴드수익률을 추가하여 학습을 진행하였다.

**4.2. 로지스틱 회귀모형을 위한 하이퍼파라미터 지정**

로지스틱 회귀모형을 학습하기 이전에 우선 하이퍼파라미터를 지정해주었다. 여기에서 사용한 하이퍼파라미터는 C, max\_iter, penalty, solver이다. 각가의 변수를 왜 사용하였고, 어떤 역할을 하는지 살펴보도록 하겠다.

C는 규제 강도의 역수이다. 규제는 과적합을 방지하기 위해 손실 함수에 패널티 항을 추가하는 것이다. 이 패널티 항은 모델의 계수 크기를 포함하며, 이 크기는 규제 강도에 의해서 제어된다. 즉, C의 값이 작을수록 강력한 규제를 받는 것이다. 우리가 만들 모델에 C가 필요한 이유는 다음과 같다. 기존의 종속변수가 8:2인 데이터를 오버샘플링을 통하여 5:5의 비율로 맞춰주었다. 그렇기 때문에 과적합 문제가 발생하기 쉽다. 실제 그냥 진행한 실험에서도 과적합이 발생되었다.

max\_iters는 최대 반복 횟수이다. 로지스틱 회귀모형은 최적의 알고리즘을 찾기 위해 많은 반복을 해야할 수도 있다. 최대 반복이 너무 낮게 설정되면 충분한 시간동안 최적의 해를 찾을 수 없고, 반대로 너무 많이 학습한다면 과적합이 발생할 수 있다.   
즉, 적정 max\_iter을 지정해야 오버피팅과 언더피팅을 예방하고 최적의 해를 찾을 수 있다.

penalty = ‘l1'은 규제화에 사용되는 norm을 지정하는 데 사용된다. 그 중에서 우리는 L1 즉 라쏘 회귀를 사용할 것이다. 라쏘 회귀를 사용하는 이유는 덜 중요한 특성의 계수를 정확하게 0으로 밀어 넣어 효과적으로 특성을 선택을 수행할 수 있다. 이를 통해 불필요한 변수는 어떤 것인지 그리고 어떤 변수가 공모주 시초가 예측에 영향을 많이 끼치는 지 파악할 수 있고, 이번 모델을 만드는데 가장 중요한 하이퍼파라미터이다.

solver = ’liblinear’은 최적화 문제에서 사용하는 알고리즘으로 로지스틱 회귀애서 손실 함수를 최소화하기 위해서 최적의 계수를 찾기 위해서 사용된다. 그중 liblinear로 지정한 것은 작은 데이터셋에 적합한 알고리즘이기 때문이다. 우리는 약 800개라는 작은 데이터셋을 사용하기에 liblinear가 적합하다. 그리고 L1과 함께 사용할 수 있다.

다시 정리하자면, 위의 4가지 하이퍼파라미터 중에서 가장 중요한 것은 penalty = ‘l1‘이다.   
그 이유는 우리는 새롭게 추가한 변수가 얼마나 공모주 시초가 예측에 많은 영향을 끼치는지 알아보기 위함이기 때문이다.

**4.3. 최적의 하이퍼파라미터를 찾기 위한 GridSearch**

우리는 최적의 하이퍼파라미터 값을 찾기 위해서 그리드서치를 사용하였다. 그리드 서치는 모든 가능한 하이퍼파라미터 조합을 시도하여 최상의 성능을 내는 하이퍼파라미터를 찾는 기법이다. 그리드 서치 이외에도 최적의 하이퍼파라미터를 찾기 위한 방법으로 랜덤서치가 있지만 이는 조합이 복잡할 때 사용하면 유용하다. 하지만 모든 조합을 고려하는게 아니니 성능이 저하될 수 있다. 우리가 진행하는 모델의 하이퍼파라미터는 조합이 복잡하지 않기 때문에 그리드 서치를 통해서 모든 조합의 경우를 수를 계산하여 최적의 하이퍼파라미터의 값을 찾을 것이다.

그리드 서치를 진행할 때 우리는 위에서 소개한 4가지 하이퍼파라미터의 조합을 살펴볼 것이다. 그리고 Cross Validation을 5로 지정하여 데이터셋이 5개의 하위집합으로 나누어지고, 각 하이퍼파라미터 조합은 각 하위 집합에 대해 훈련과 검증을 이루어서 결과의 편향을 줄이고 보다 더 정확한 값을 얻을 수 있다. 그리고 n\_jobs를 5로 지정하여서 5개의 하위 집합을 동시에 작업하여 그리드서치의 속도를 높였다.

**4.4. 데이터 Scaler과 split**

StandardScaler를 통하여 평균을 0, 표준편차를 1로 변경하여 데이터를 정규화한다. 즉, 특성의 값이 정규분포를 따르도록 하는 것이다. 이는 각 특성의 스케일이 비율 혹은 백만원 단위의 수치값으로 차이가 많이 나기 때문에 사용한다. 이를 통해서 수치를 안정화해준다.

split는 0.2로 지정하여준다. 전체 데이터셋의 양이 적기 때문에 학습 데이터를 8로 많이 지정하여 학습을 진행하였다.

**4.5. CART-LR**

그리드서치로 최적의 하이퍼파라미터를 탐색한 결과는 다음과 같다. C: 100, max\_iter: 600 나머지 penalty와 solver은 각각 l1과 liblinear로 고정이다. 이를 통하여 학습한 결과는 다음과 같다.

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

accuracy는 0.8 그중에서 종속변수가 1일때의 recall은 0.84로 괜찮은 모델 성능이 나온 것을 확인할 수 있다. 유의미한 모델 결과를 나타냈기 때문이 이 모델을 통해서 살펴본 각 feature의 중요도를 살펴보면 다음과 같이 나온다.

텍스트, 스크린샷, 디스플레이, 번호이(가) 표시된 사진

자동 생성된 설명

위의 결과를 보았을 때 청약 경쟁률이 월등히 높은 중요도를 가지고 있는 것을 확인할 수 있다.   
참고로 이는 상대적인 중요도이지 절대적 중요도는 아니다. 그리고 청약 경쟁률에 이어서 공모가, 의무보유확약이 중요한 독립변수로 자리 잡은 것을 확인할 수 있다. 그 외에도 최근 “따” 비율, 전문투자자(비율)도 각각 10번째, 11번째로 중요한 변수로 지정되어 있는 것을 확인할 수 있다.

**4.6. C5.0-LR**

그리드서치로 최적의 하이퍼파라미터를 탐색한 결과는 다음과 같다. C: 1,000, max\_iter: 1000 나머지 penalty와 solver은 각각 l1과 liblinear로 고정이다. 이를 통하여 학습한 결과는 다음과 같다.

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

결과는 CART-LR과 동일하게 나왔으며 유의미한 모델로 판별된다. 따라서 이를 통해서도 유의미한 변수들이 무엇인지 살펴보았다.

텍스트, 스크린샷, 디스플레이, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

이 역시 청약 경쟁률이 가장 월등히 높은 중요도를 내포하고 있다. 뒤이어서 공모가, 의무보유확약이 중요한 독립변수로 자리잡았다. 추가 변수를 살펴보자면 여기에서는 전문투자자(비율), 밴드수익률, 최근 “따” 비율이 각각 7번째, 9번째, 13번째로 중요한 변수임을 확인할 수 있었다.

**4.7. 시사점 정리**

결론적으로, 기본 15개 변수 중 청약경쟁률, 공모가, 의무보유확약이 Top3로 공모주 시초가 예측에 영향을 미쳤고, 그 중에서 청약 경쟁률이 가장 큰 영향을 미치고 있다. 그 외에도 기본변수들이 꽤 많은 영향을 미치는 것을 확인할 수 있다.

그렇다고 추가한 변수가 영향이 없는 것은 아니다. ‘ROE’를 제외한 ‘전문투자자비율’, ‘밴드수익률’, ‘최근 “따” 비율’이 모두 유의미한 영향을 미치고 있는 것을 확인할 수 있었다. 우선은 기본 변수 중 Top3 변수와 추가변수가 어느 시점이 되었을 때 유의미한 값을 나타내는지 살펴보도록 하겠다.

**5. 변수의 유의미한 지점 도출**

**5.1 기본 변수 중 Top3 변수의 유의미한 지점**

해당 작업은 선정된 6개의 변수가 어느 지점이 되어야지 공모주 시초가 100% 예측에 유의미하게 영향을 미치는지 그 지점을 확인하는 것이다. 로지스틱 회귀모델을 통해서 회귀선을 그려서 종속변수 1의 회귀선과 0의 회귀선이 교차하는 지점을 유의미한 지점으로 해석된다.

**5.1.1 로지스틱 회귀곡선의 Top3 변수 회귀선**

|  |
| --- |
| 텍스트, 스크린샷, 라인, 그래프이(가) 표시된 사진  자동 생성된 설명 |
| 텍스트, 스크린샷, 도표, 그래프이(가) 표시된 사진  자동 생성된 설명 |
| 텍스트, 스크린샷, 라인, 그래프이(가) 표시된 사진  자동 생성된 설명 |

이 그래프에서는 청약 경쟁률은 약 1,000 이상, 공모가의 경우에는 약 13,500 이하, 의무보유확약은 약 0.15 이상이 되어야 공모주 시초가 100% 예측에 유용한 의미를 가진다는 것을 알 수 있다.

**5.1.2 로지스틱 회귀곡선의 추가 변수 회귀선**

|  |
| --- |
| 텍스트, 스크린샷, 라인, 그래프이(가) 표시된 사진  자동 생성된 설명 |
| 텍스트, 스크린샷, 그래프, 라인이(가) 표시된 사진  자동 생성된 설명 |
| 텍스트, 스크린샷, 라인, 그래프이(가) 표시된 사진  자동 생성된 설명 |

이 그래프에서는 ‘최근 “따” 비율’은 약 0.225 이상, ‘밴드수익률’은 약 0.95 이상이 되어야 공모주 시초가 100% 예측에 유용한 의미를 가진다는 것을 알 수 있다. 하지만 ‘전문투자자(비율)’은 가설 예측과 정반대의 결과로 나왔기 때문에 시초가 예측에 무의미한 변수라는 것을 알 수 있었다.

**6. 최종 모델 구현**

**6.1 최종 모델 구현 (인공신경망)**

기본 15개의 변수 중 가장 영향을 적게 미치는 ‘자본금(백만원)’, ‘순이익(백만원)’을 제외한 기본 변수 13개와 추가 변수인 ‘최근 “따” 비율’, ‘밴드수익률’을 추가해 총 15개의 변수로 최종 모델을 학습을 진행하였다.

|  |  |
| --- | --- |
| **변인** | **변수명** |
| X1 | 상장 시기(상장월) |
| X2 | 최근 "따" 비율 |
| X3 | 공모가 |
| X4 | 밴드수익률(공모가/공모가\_상한) |
| X5 | 기관경쟁률 |
| X6 | 청약경쟁률 |
| X7 | 의무보유확약 |
| X8 | 대표주간사 |
| X9 | IPO 주간사 순위 |
| X10 | 구주매출(비율) |
| X11 | 매출액(백만원) |
| X12 | 총 공모주식 수 |
| X13 | 우리사주조합(비율) |
| X14 | 공모후 상장주식수(주식수) |
| X15 | 유통가능주식합계(비율) |
| Y | 종속변수 |

**6.2 하이퍼파라미터 선정**

하이퍼파라미터는 layer의 층을 나타내는 hidden\_layer\_sizes와 learning\_rate\_init, max\_iter, Early Stopping, alpha를 사용하였다.

여기에서 learning\_rate는 너무 높게 설정되면 학습이 빨라서 불안정해져 정확한 최적의 해를 찾을 수 없게 되고 너무 낮게 되면 지역 최적점에 갇히게 될 수 있다. 그래서 그리드 서치를 통해서 적정 학습률을 선택해줄 것이다.

Early Stopping은 검증 점수가 향상되지 않을 때 훈련이 자동으로 중지가 된다. 이렇게 되면 과적합이 방지가 된다. 학습 데이터셋이 적은 이번 실험에서는 반드시 필요한 하이퍼파라미터이다.

alpha는 정규화의 강도를 조절하는 매개변수이다. 모델의 복잡도에 패널티를 부여하여 모델이 과적합 되는 것을 방지해준다. 값이 클수록 더 강한 정규화가 적용되고, 작을수록 약한 정규화가 적용된다. 이 역시 마찬가지로 학습데이터가 적은 이번 프로젝트에서 반드시 필요한 하이퍼파라미터이다. 참고로 해당 하이퍼파라미터는 0.0001로 고정할 것이다. 값이 여러 실험결과 과적합 혹은 과소적합이 되지 않는 값으로 보인다.

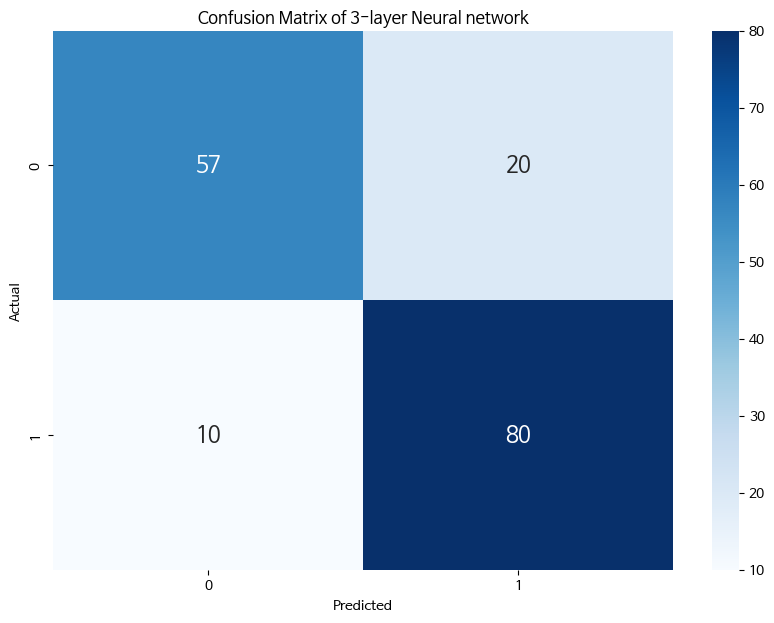
그리드 서치로 최적의 하이퍼파라미터를 선정한 결과는 다음과 같다.

|  |  |  |
| --- | --- | --- |
| **하이퍼파라미터** | **3계층** | **4계층** |
| max\_iter | 100 | 400 |
| learning\_rate\_init | 0.1 | 0.1 |
| alpha | 0.0001 | 0.0001 |
| hidden\_layer\_sizes | 15 | 15,15 |
| early\_stopping | True | True |

**6.3 최종 모델 선정**

|  |  |
| --- | --- |
| **인공 신경망 3계층** | **인공 신경망 4계층** |
|  |  |

최종 모델은 인공신경망 3계층이다. 4계층에 비해서 3계층이 accuray가 0.82로 더 뛰어나고 종속변수 1일 때 recall은 0.89로 동일하다. 전체적인 측면에서는 인공신경망 3계층이 보다 더 적합한 모델인 것을 확인하였다.



위의 이미지는 인공신경망 3계층의 최종결과표이다. 공모주 시초가 100%인 종속변수 1이 90개중 80개를 맞추었다.

**7. 결론 및 한계점**

본 보고서에서는 코스닥시장에서의 공모주 시초가 예측에 대해 알아보았다. 이를 위해 공모주 시초가 예측을 위해 주로 사용하는 변수 15개로 상장 시기(상장월), 공모가, 기관경쟁률, 청약경쟁률, 의무보유확약, 대표주간사, IPO 주간사 순위, 총 공모주식 수, 구주매출(비율), 매출액(백만원), 순이익(백만원), 자본금(백만원), 우리사주조합(비율), 공모후 상장주식수(주식수), 유통가능주식합계(비율)를 사용하였다.  
이 중에서 로지스틱 회귀모형으로 분석한 결과 가장 영향이 적었던 2가지 변수인 ‘순이익(백만원)’, ‘자본금(백만원’)을 제외한 13개를 기본 변수로 선정하였다. 그리고 추가 변수 중 가장 유의미한 ‘최근 “따” 비율’, ‘밴드수익률(공모가/공모가\_상한)’을 추가해 기본변수 13개와 추가변수 2개 총 15개 변수로 공모주 시초가 100% 예측 최종 모델을 만들었다.

여기에서 우리는 전체 변수 15개 중에서 가장 의미가 있었던 변수는 **‘청약 경쟁률’, ‘공모가’, ‘의무보유확약’**이 이었고, 청약 경쟁률은 약 1,000 이상, 공모가의 경우에는 13,500원 이하, 의무보유확약은 0.15 이상이 되면 공모주 시초가 100% 예측에 유용한 의미를 가진다는 것을 확인할 수 있었다.

또한, 가설로 정했던 추가 변수인 **‘최근 “따” 비율’, ‘밴드수익률(공모가/공모가\_상한)’**가 공모주 시초가 100% 예측에 유의미한 영향을 끼친다는 것이다. 이러한 결과를 도출하기 위해서 3번의 검수 과정을 거쳤다.

첫 번째로는, 의사결정나무의 CART와 C5.0으로 추가 변수 4가지 중에서 ‘ROE’를 제외한 위의 세가지 변수가 유의미하다는 것을 확인하였다.   
두 번째로는, 로지스틱 회귀모형을 통해서 공모주 시초가 100% 예측에 영향을 끼치는 변수의 중요도를 확인함으로써 위의 세가지 변수가 모두 유의미하다는 것을 확인하였다.   
마지막 세 번째로, 로지스틱 회귀모형의 회귀선에서 ‘최근 “따” 비율’, ‘밴드수익률(공모가/공모가\_상한)’은 공모주 시초가 예측 100% 달성에 유의미한 결과를 나타냈고, ‘전문투자자(비율)’은 가설과 정반대의 그래프를 형성하므로써 제외되었다.

이로써 4가지 가설 중, 재무건전성 파악을 통해서 저평가된 종목을 찾는 ‘ROE’와 전문 투자자의 비율이 높으면 일반 투자자들에게 기대심리를 주어 더욱 활발한 투자를 불러 일으킬 수 있을 거라고 예측했던 ‘전문투자자(비율)’은 시초가 예측에 의미가 없었고, ‘최근 “따” 비율’을 통하여 최근 주식 시장의 동향을 반영하는 것과 ‘밴드수익률’을 통하여 공모가격이 희망 공모가 상한에 얼마나 가까운지 측정하여 투자자들의 기대심리를 반영하는 가설이 맞았음을 확인할 수 있었다.

한계점은, 1장에서 설명한대로 우리는 가장 최근의 자료를 사용하기 위해 2015년 6월 15일부터 현재 2023년 5월까지의 8년치 코스닥 공모주 데이터 528개를 웹 크롤링하여 가져왔다. 하지만 데이터의 범위가 ‘금융당국이 전일 종가 대비 가격제한폭을 ±30%로 확대한 지점’을 기간으로 설정하였기 때문에 데이터의 양이 많지 않았다.

그리고 2020년엔 ‘코로나 펜데믹’으로 인해 주가의 대폭락이 있었다. 이후 많은 국가들이 코로나 대응을 위해 경제적인 지원 및 정책조치를 시행해 주식시장에 긍정적인 신호로 작용하여 주가가 급속도로 상승하기도 하였지만, 주식시장이 일반적이지 않았다는 점이 한계점으로 들 수 있다.

**『참고문헌』**

[1] Using Machine Learning Algorithm for Forecasting Returns on IPO firms, 최연지, 2022

[2] A study on the initial public offering subscription rate and stock price behavior after IPO : Focused on kosdaq market, 백삼열, 2019

[3] 투자주체별 청약경쟁률과 공모주(IPO)수익률에 관한 연구, 김종근, 2016

[4] Using AI to develop forecasting model in IPO market, 조득환 외 3인, 2020

[6] <https://www.thebell.co.kr/free/content/leaguetable.asp?kind=dcm&lcode=03&year=2014>

[7] http://www.38.co.kr/

[8] http://ipostock.co.kr/main/main.asp

[9] https://www.pstock.co.kr/2005pstock/main.asp

**[5조] 팀 프로젝트 구성원 별 참여사항**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 조원 | 학번 | 대면 회의(5/31) | 카톡 회의 | 역할 |
| 고도환 | 5434535 | O | O | 가설설정, 모델링, 결과해석, 보고서 작성 |
| 전시영 | 4601263 | O | O | 가설자료조사, 결과해석, 보고서 수정 및 보완 |
| 정한별 | 5533017 | O | O | 데이터수집 및 분석, 발표 |
| 김우준 | 5605627 | O | O | 가설자료조사, PPT만들기 |
| 풍티하이하 | 5731440 | X | X | 미참여 |