서론

IPO(Initial Public Offering)는 기업공개라고 하는데, 일정 규모의 기업이 상장절차 등을 밟기 위해 외부 투자자가 공개적으로 주식을 살 수 있도록 기업이 자사의 주식과 경영 내역을 시장에 공개하는 것이다. IPO는 기업이 주식을 상장하는 방법 중에서 가장 많이 활용되는 방법인데 상당히 절차가 복잡하고 까다롭다.

폰트, 원, 봄, 자연이(가) 표시된 사진

자동 생성된 설명

<그림1> IPO 절차 (신한투자증권)

<그림1> IPO 절차는 1)비상장기업은 먼저 증권선물거래소에 상장신청을 한다. 증권선물거래소에서는 상장신청을 한 기업들을 대상으로 심사를 하게 된다. 2) 심사기준을 충족하게 되면 승인을 내주고, 기업은 IPO를 위한 본격적인 실물 작업에 들어간다. 3) 대표주관사를 선정하고 주관사와 함께 증권신고서(공모방법, 공모가 결정방법, IPO를 하는 목적, 기업의 재무제표, 주주현황, 기업의 사업현황 등 포함)를 작성한다. 4)기관투자자들을 대상으로 수요예측을 진행한다. 5) 수요예측이 끝나고 나면 공모가를 결정하게 된다. 공모가가 결정되면 기업의 시가총액이 결정되고, 공모주 청약이 시작된다. 6)공모주 청약이 끝나고 일정시간 후에 정식으로 주식시장에 상장을 한다.

IPO의 장단점이 존재한다. 장점으로는 첫 번째로 IPO는 주주들로부터 대규모 자금조달이 가능하다. 두 번째로는 기업이 상장을 하게 되면 기업의 인지도가 높아지고 평판이 좋아진다. 세 번째로는 상장된 기업이라면 시장에서 검증되었다고 판단하기 때문에 신용도가 상승하고 그로인해 저금리로 차입이 가능하다. 마지막으로 기존 투자자들은 본인의 투자자금을 회수할 수 있다.

하지만 단점도 존재한다. 첫번째로는 경영권 분산이 있다. 상장주식수와 투자자 수의 증가로 경영권이 분산되어 침해를 받을 수 있다. 두 번째로는 투자자들과의 의사소통 어려움이 있다.주주들은 향후 실적이나 전망에 관심이 많고 그 기대에 부응해주어야 하기 때문에 의사소통 강화필요성이 있다. 마지막으로 IPO를 준비하는 기간(6개월~1년)동안 많은 비용이 발생한다.

이런 IPO에 투자자들이 관심을 많이 가지는 이유는 단기 수익률이다. 가격 변동성이 큰 일반 주식과 달리 상장 당일 큰 가격변동이 일어나기 때문에 단기간의 수익을 바라고 투자하는 투자자들이 많다. 현재 신규 상장 종목의 경우 공모가격의 90~200% 사이에서 호가를 접수해 결정된 시가를 신규 상장일 기준가격으로 사용하고 가격 제한폭은 기존 상장 종목과 마찬가지로 하한 -30%, 상한 +30%로 그어진다. 시초가가 공모가의 2배로 형성만 되어도 단기간에 큰 수익을 낼 수 있기 때문에 투자자들에게 관심이 뜨거웠다.

텍스트, 폰트, 번호, 스크린샷이(가) 표시된 사진

자동 생성된 설명

<그림2> 주가제한폭 변경안

하지만 6월말부터 신규 상장 당일 주식의 기준가격 결정 방법이 달라진다 <그림2>. 신규상장일 당일 신속하게 균형가격을 찾도록 하기 위한 취지로 미국, 일본, 중국 등 해외 사례를 참고해 신규상장종목의 기준가는 별도의 절차 없이 공모가격 그대로 결정되며 가격제한폭이 공모가의 60~400%로 확대된다. 즉 시초가라는 개념이 사라진다. 고수익을 노린 투자자들이 공모주 시장으로 몰려들면서 청약경쟁이 더욱 치열해질 전망이다. 또, 변동폭이 커진만큼 높은 수익을 낼 수도 있지만 상황이 좋지 않은 경우 60%의 손실도 발생할 수 있다는 점을 봤을 때 공모주 투자 난도가 높아지고 투자 위험이 커져 투자자들이 공모주 투자에 신중을 기할 것이다.

연구목적 및 필요성

정보를 얻는데 한계가 있는 개인투자자들은 심각한 역선택의 문제에 직면해 저평가된 공모주에 비해 상대적으로 고평가된 공모주를 많이 배정받기 때문에 정(+)의 초과수익률을 달성할 수 없다는 선행연구(Rock 1986, Beaty and Ritter 1986, Michaely and Shaw 1994, Amihud, Hauser and Kirsh 2003)가 있기 때문에 투자 의사결정에 어려움을 겪을 것이다. 6월 말 60~400%로 확대될 뿐더러 거래량이 한정되어있기 때문에 현재보다 큰 수익을 낼 수 있지만 큰 손실로 이어질 수 있고 상장 첫날 관심이 가장 많기 때문에 가격변동성이 클 것이다. 개인투자자의 청약 경쟁이 과열될수록 역선택 문제로인한 정보가 없는 개인투자자의 기대 투자 손실이 커지기 때문에 때문에 앞으로는 단순히 접근하기보다는 비재무적정보와 재무적정보를 잘 판단해 공모주 투자 전략을 세워야 한다. 이를 통해 본 연구에서는 그에 대비해 정보의 비대칭성의 피해를 최소화 하여 높은 손실을 줄이고 투자 의사결정에 도움을 주고자 최적의 모델을 구현 하고자한다. 공모가를 기준으로 등락할지 하락할지 예측하는 모델을 구현한다. 머신러닝 기반의 분류모델 이용해 ‘전자공시시스템’, ‘38커뮤니케이션’에서 제공되는 축적된 기업 데이터를 수집해 수익여부(label)에 영향을 끼치는 요소들을 찾고 알고리즘들의 성능을 비교 후 최적의 모델로 6월말 새로운 방법으로 적용된 데이터들에 적용하여 예측할 수 있도록 한다.

연구의 구성은 데이터, 탐색적데이터분석, 전처리, 모델링, 결과, 결론을 순서로 진행하였다.

연구 데이터

‘시초가 예측데이터’와 ‘38커뮤니케이션’, ‘전자공시시스템’에 축적된 데이터들을 크롤링하여 추가 수집해서 총539개의 기업을 분석하였다. 변수로는 ‘기업명’, ‘공모가(원)’, ‘ 기관청약경쟁률’, ‘일반청약경쟁률’, ‘상장주식수’, ‘매출액(백만원)’, ‘순이익(백만원)’, ‘자본금(백만원)’, ‘시장구분’, ‘의무보유확약비율’, ‘구주매출비율(%)’, ‘상장일’, ‘ 최대주주지분율(%)’, ‘주간사 명’, ‘유통가능물량(%)’, ‘최대희망공모가액차(%)’ ,’수익여부’, ‘공모금액(백만원)’ ,’종가’ 총 19개의 변수를 사용하였다. 종속변수는 ‘수익여부’로 하였다. ‘기업명’ , ‘시장구분’, ‘상장일’, ‘주간사명’ 을 제외한 모든 변수는 수치형 변수이다.

탐색적 데이터 분석

전처리

통계적 검정

범주형 독립변수에 따른 수익의 차이가 있는지 알고 싶지만 종속변수인 ‘수익여부’가 범주형이기 때문에 ‘첫날손익률’이라는 파생변수를 만들어 통계적 검정을 진행하였다.

‘상장분기’, ‘주간사규모’, ‘유가증권여부’ 3가지 범주형 변수에 대해서 ‘첫날손익률’이 차이가 있는지 통계적 검정을 하였다. 통계적 검정방법으로는 모수적 검정 방법과 비모수적 검정 방법이 있다. 모수적 검정 방법은 3가지 가정이 모두 만족하여야 사용할 수 있다.

1) 정규성 : 각 집단에서 표본은 정규성을 가진다.

2) 등분산성 : 각 집단의 분산은 서로 동일해야 한다.

3) 독립성 : 표본은 서로 독립적이어야한다.

이 3가지 가정이 하나라도 만족하지 않으면 비모수 검정 방법을 사용하였다. 정규성 검정으로는 샤피로-윌크 검정(Shapiro-Wilk test), 등분산성 검정으로는 바틀렛 검정(Bartlett test), 레빈 검정(Levene test)를 사용하였다. 먼저 ‘상장년도’ 변수는 2014~2022년까지의 14개의 범주로 구성되어져있다. 각 집단의 표본 크기가 33~86으로 형성되어있다. 표본 크기가 작고 샤피로-윌크 검정(Shapiro-Wilk test)을 통해 정규성을 검정하였을 때, p-value < 0.05로 정규성을 만족하지 못하였다. 그래서 비모수적 검정 방법인 크루스칼-왈리스(Kruskal-Wallis) 검정을 진행하였다. p-value < 0.05로 집단사이의 유의미한 차이가 있다고 나타났다. 두 번째로 ‘상장분기’ 변수는 1~4분기까지의 4개의 범주로 구성되어져 있다. 각 집단의 표본 크기가 62~228로 형성되어져 있고 ‘상장변수’ 샤피로-윌크 검정(Shapiro-Wilk test)을 통해 정규성을 검정하였을 때, p-value < 0.05로 정규성을 만족하지 못하였다. 그래서 비모수적 검정 방법인 크루스칼-왈리스(Kruskal-Wallis) 검정을 진행하였다. p-value < 0.05로 집단사이의 유의미한 차이가 있다고 나타났다. 세 번째로 ‘주간사규모’ 변수는 ‘대형’, ‘대형아님’으로 2가지 범주로 구성되어져있다. 각 집단의 표본 크기는 451,88로 형성되어있고 샤피로-윌크 검정(Shapiro-Wilk test)을 통해 정규성을 검정하였을 때, p-value < 0.05로 정규성을 만족하지 못하였다. 그래서 비모수적 검정 방법인 맨-휘트니 U 검정(Mann-Whitney U test)을 진행하였다. p-value > 0.2로 1%로 유의수준에서도 귀무가설을 채택하고 집단사이의 유의미한 차이가 없다고 나타났다. 마지막으로 ‘유가증권여부’ 변수는 ‘유가’, ‘유가아님’으로 2가지 범주로 구성되어져있다. 각 집단의 표본 크기는 63,476으로 형성되어있고 샤피로-윌크 검정(Shapiro-Wilk test)을 통해 정규성을 검정하였을 때, p-value < 0.05로 정규성을 만족하지 못하였다. 그래서 비모수적 검정 방법인 맨-휘트니 U 검정(Mann-Whitney U test)을 진행하였다. p-value < 0.05로 집단사이의 유의미한 차이가 나타났다.

모델링

* train , test 분리

데이터수가 539개로 적기 때문에 train , test 셋을 6:4로 나누어서 진행하였다

* 정규화

독립변수들은 수치 단위 차이가 커 ‘상장분기’ , ‘주간사규모’, ‘유가증권여부’ 3가지 범주형 변수를 제외하고 수치형 변수에 대해 정규화 작업을 진행해주었다.

정규화는 Python의 MinMaxScaler을 사용하였다. 정규화 식은 (1)과 같다.

1. Extreme Gradient Boost

Gradient Boosting은 단순하고 약한 모델 세트의 추정치의 앙상블을 결합하여 대상변수를 정확하게 예측하려는 시도하는 지도 학습 알고리즘이다.

Extreme Gradient Boost의 일반적인 구조는 Gradient Boosting의 고급버전이다. 즉, 잔차로부터 훈련한 트리를 추가하여 약한 학습기를 강력한 학습기로 바꾼다. Extreme Gradient Boost가 Gradinet Boosting보다 선호되는 이유는 더 좋은 성능을 내고 더 빠르기 때문이다.Extreme Gradient Boost는 기본 학습기를 의사결정나무로 하며 Gradient Boosting과 같이 Gradien(잔차)를 이용하여 이전 모형의 약점을 보완하는 방식으로 학습 하는 알고리즘이다. 또, Gradient Tree Boosting에 과적합 방지를 위한 파라미터가 추가된 알고리즘이다.

GridSearchCV를 통해 최적의 하이퍼 파라미터를 찾아주었다.

colsample\_bytree = 0.8

gamma = 0

learning\_rate = 0.15

max\_depth = 10

n\_estimators = 100

파라미터로 XGBClassifier을 학습하였다.

정확도 : 0.81

정밀도 : 0.84

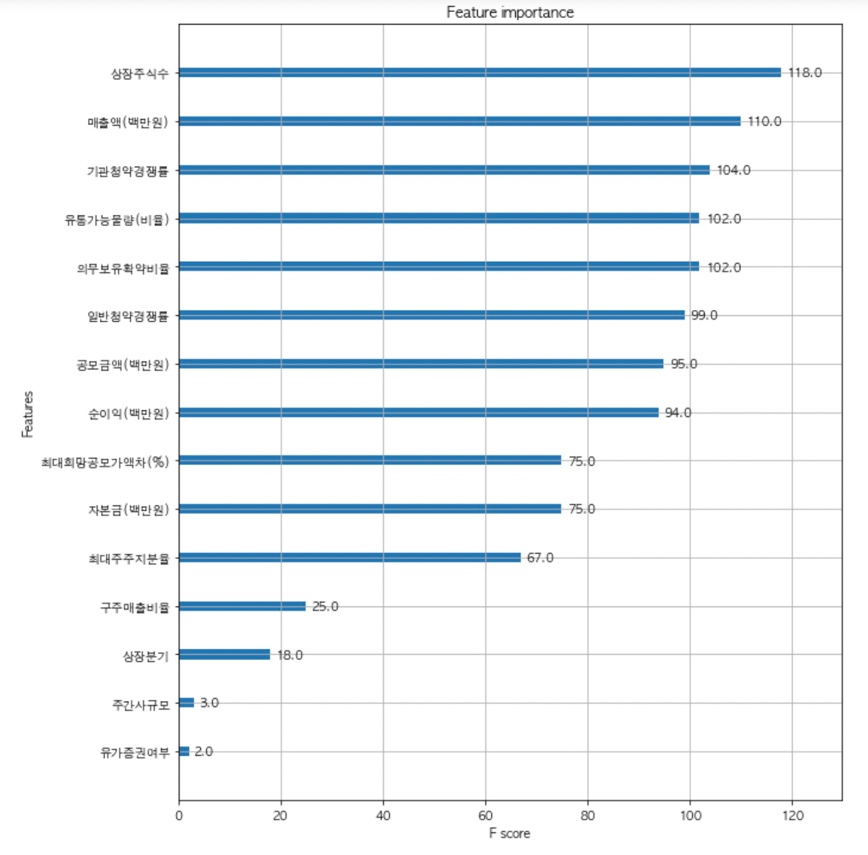
재현율 : 0.91

F1 – score : 0.88

AUC : 0.72

텍스트, 폰트, 화이트, 타이포그래피이(가) 표시된 사진

자동 생성된 설명



다른 모델과 달리 F score 를 기준으로 Extreme Gradient Boosting의 변수중요도 상위 10개를 나타내었다. F score는 Tree를 분할할 시 해당 feature가 자주 사용되었는가를 나타내는 지표이다. F score가 높을수록 Tree분할에 더 잘 사용되는 것이다.

상장되는 주식의 수가 가장 중요한 것으로 나타났다. 또, ‘상장주식수’, ‘기관청약경쟁률’, ‘유통가능물량(비율)’, ‘의무보유확약비율’ 등 비재무정보가 상위권에 있는 것을 확인할 수 있다.

1. Random Forest

Random Forest는 기존 배깅의 이점을 살리고 변수를 랜덤으로 선택하는 과정을 추가함으로써 개별 나무들의 상관성을 주여서 예측력을 향상한 앙상블 모형이다.

Random Forest는 기본적으로 배깅을 사용하기 때문에 배깅의 장점인 편의(Bias)를 유지하면서 분산을 낮추는 효과를 Random Forest 또한 그대로 이어받는다.

(2)에서 Random Forest가 배깅을 사용하는 과정을 보여준다.

도표, 텍스트, 라인, 폰트이(가) 표시된 사진

자동 생성된 설명

Bootstrap Sampling을 통해 얻은 샘플 데이터를 이용하여 개별 나무들을 학습시키고 주어진 입력데이터 를 B개의 개별 나무에 넣어서 얻은 출력값들을 집계하여 예측값을 얻게 된다. Random Forest에서 개별 나무간의 상관성이 작아지면 일반화의 오류가 작아지는 것이 논문에서 증명이 되었고, 즉 상관성을 줄이면 예측력이 좋아진다는 것이다. 종합하면 Random Forest는 개별 나무를 성장할 때 분리마다 랜덤으로 변수 후보를 선택하고 이를 통해 개별 나무 간의 상관성을 줄여준다.

Random Forest 또한 GridSearchCV를 통해 최적의 하이퍼 파라미터를 찾아주었다.

n\_estimators = 10

max\_depth = 6

min\_samples\_leaf = 8

min\_samples\_split = 8

random\_state = 0

n\_jobs = -1

파라미터로 RandomClassifier을 학습하였다.

텍스트, 폰트, 화이트, 디자인이(가) 표시된 사진

자동 생성된 설명

정확도 : 0.79

정밀도 : 0.82

재현율 : 0.91

F1 – score : 0.86

AUC : 0.68

*텍스트, 스크린샷, 다채로움, 라인이(가) 표시된 사진

자동 생성된 설명*

Random Forest의 변수중요도를 나타내었다. Extreme Gradient Boosting과 달리 ‘일반청약경쟁률’이

중요도가 높은 것으로 나타났다.

1. Extra Tree

Extra Tree는 극도로 무작위화된 기계 학습 방법이다. 데이터 샘플 수와 특성 설정까지 랜덤으로 이루어지며 Random Forest와 동일한 원리를 이용하기 때문에 많은 특성을 공유한다. Random Forest에 비해 속도가 빠르고 성능도 미세하게 높다. Bootstrap Sampling을 사용하지 않고 전체 특성 중 일부를 랜덤하게 선택해 노드 분할에 사용한다. 무작위 분할 중 가장 좋은 것을 분할 규칙으로 선택한다.

Extra Tree의 파라미터를 GridSearchCV 통해 찾아주었다.

n\_estimators = 25,

max\_depth = 4

random\_state = 0

n\_jobs = -1

파라미터로 ExtraTreesClassifier을 학습시켰다.

텍스트, 폰트, 화이트, 디자인이(가) 표시된 사진

자동 생성된 설명

정확도 : 0.74

정밀도 : 0.74

재현율 : 0.98

F1 – score : 0.85

AUC : 0.53

텍스트, 소프트웨어, 다채로움, 번호이(가) 표시된 사진

자동 생성된 설명

Extra Tree의 변수중요도를 나타내었다. 청약경쟁률이 가장 중요도가 높은 것으로 나타났다.

또, 다른 모델과 달리 ‘최대희망공모가액차(%)’ 변수도 중요도가 상위권인 것으로 나타났다.

1. SupportVectorMachine

SupportVectorMachine의 파라미터를 GridSearchCV 통해 찾아주었다.

C = 100

gamma = 1

파라미터로 SupportVectorMachine을 학습시켰다.

텍스트, 폰트, 화이트, 디자인이(가) 표시된 사진

자동 생성된 설명

정확도 : 0.76

정밀도 : 0.82

재현율 : 0.87

F1 – score : 0.84

AUC : 0.67

결론

‘수익여부’ 종속변수는 0 : 손실, 1 : 수익이다.

중요하게 봐야할 것은 수익성 있는 상장주식을 예측하는 것이 목표이기 때문에 1을 예측한 것에서 실제로 1을 보는 정밀도를 중요하게 보았다. Extreme Gradient Boosting이 0.84로 가장 뛰어난 성능을 보였다. 그 뒤로 SupportVectorMachine과 RandomForest과 뒤를 이었고 ExtraTree가 가장 좋지 않은 성능을 보였다. SupportVectorMachine을 제외한 트리형 모델 3개에서 공통적으로 청약율이 변수 중요도에서 상위에 있는 것을 확인하였다. 또, 재무적정보 보다 비재무적 정보가 더 중요한 변수인 것을 확인할 수 있었다. 그 중에서도, 6월 말 제도가 바뀌면 기관의 수요예측기간이 연장 된다. 또, 주관사가 기관 주금납입 능력을 확인 후 배정하고 공모가 미기재 수요 예측기관은 공모주를 배정받지 못하는 등 기관의 허수적 청약을 방지하기 위해 규제가 강화가 될 예정이므로 기관청약율이 더 중요해질 전망이다. 그리고 2015년 6월 15일 ±15%로 묶여 있던 가격제한폭이 ±30%로 확대 되었을 때 영향을 미쳐 공모수익률과 개인 청약경쟁률의 유의한 정(+)의 상관관계는 낮아졌으며, 공모수익률과 기관 청약경쟁률의 유의한 정(+)의 상관관계는 높아졌다(백삼열). 이 선행연구로 보았을 때, 6월말 가격제한폭이 확대되면 일반청약경쟁률보다 기관청약경쟁률을 더 중요시하게 봐야하는 것에 뒷받침하는 근거가 되었다. 의무보유확약기간별 물량도 분산배정될 예정이오니 중요도가 올라갈 것이다. 하지만 주간사규모는 유용한 정보가 되지 못하고 있다. 우리나라 IPO 시장에서 투자성과를 예측하는데 주간사 평판도는 유용한 공개정보가 되지못하고 있다는 선행연구를 보았을 때(김주환,박진우), 주간사는 유용한 정보가 아니라고 판단된다.

참고문헌

[1]공개정보와 IPO 주가성과 (김주환,박진우),2019

[2]공모주 청약경쟁률과 IPO 이후 주가행태에 관한 연구 –코스닥시장을 중심으로- ,백삼열, 2019

개인투자자의 IPO 장단기 투자성과 민재훈 2018