「2023 년 1 학기 빅데이터 경제학」

데이터 분석 프로젝트 보고서

KRX 신규 상장 주식 주가 예측

이름 고민성　　학번 20182788

Project instruments

최근 들어 신규 상장 불패 신화 관련, 뉴스가 많이 들려오고 있다. 또한, 한때 공모주 열풍이 불기도 하였다 그렇다면, 진짜로 공모주는 불패 신화일까? 항상 공모주에 대해서 공모하는 것이 이득이 되는지를 알아보기 위해 공모주의 가격을 예측하여 공모주 선택 및 공모주의 매도 시기 결정에 도움을 주고자 한다.

Project Goals

KRX 신규 상장 주식을 분석하여 어떠한 종류의 공모주를 선택하는 것이 좋은지, 언제 파는 것이 좋은지에 대한 분석 진행 및 모델을 개발하여 현명한 주식 투자를 하고자 한다.

Methods

공공데이터포털에서 신규 예탁지정 종목조회를 통해 신규상장 종목들의 특성을 확인해 보고 주식 상장의 수는 계절성을 가지는 것을 확인한다. 공공데이터포털에서 금융위원회_주식 시세정보를 수집한다. 이때, API 키를 이용하며 반복문을 통해 데이터를 축적하며 수집한다. 금융위원회_주식 시세정보를 이용하여 20~21년도 상장기업 중 300거래일 이상의 정보가 있는 161개의 데이터를 수집한다. 이때, 공모주의 이름의 '스팩'이 들어갈 경우 이는 제가 분석하고자 하는 대상과 특성이 다른 점이 존재하여 제외한다. 공모주의 불패 신화를 증명하기 위해 랜덤적으로 공모주를 선택하여 300거래일 이상 보유했을 경우의 수익을 확인하기 위하여 161개의 데이터 중 랜덤으로 50개를 추출하는 작업을 반복한다. 그리고 50개의 신규 상장 공모주를 300거래일 동안 보유하였을 때 기대되는 수익률을 추출한다. 그리고 이를 히스토그램으로 시각화한다. X축은 수익률을 나타낸다.

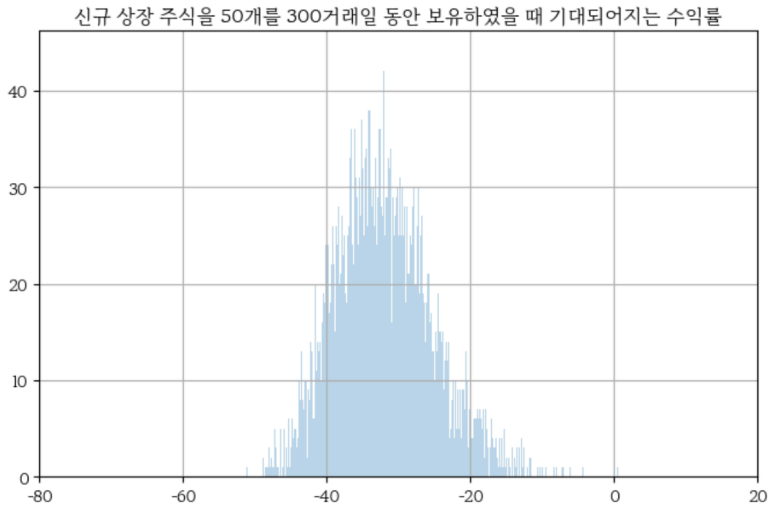

Figure 1 신규 상장 주식 50개에 대한 300거래일 수익률 히스토그램

랜덤적으로 50 개의 공모주를 선정하여 300 거래일을 보유할 경우 위의 자료와 같이 대부분 손해를 보는 것으로 나타났다. 이는 20~21 년도 당시 코로나로 인한 주식시장 여파로 분석이 되지만, 외부 요인을 신경 쓰지 않고 공모주에 투자할 경우 손해를 볼 수 있다는 것을 나타낸다. 이는 공모주가 무조건적인 불패를 나타내는 것임이 아님을 밝힌다. 그렇기에 우리는 공모주 투자를 위해 현명한 방식을 취해야 한다. 그렇기에 이 보고서에서 여러 가지 기업의 내부적, 외부 환경의 요인들을 고려하여 공모주를 선택하고 적절한 매도 시기를 선정하여 수익을 올릴 방안을 제시한다.

추가적인 분석을 진행하기 위해 대한민국 대표 기업공시 채널 KIND 에서 공모기업 현황. 신규상장기업 현황, 영업성과 추이, 일반재무 현황, 종목별 공모가 대비 주가 등락률 현황 데이터를 수집한다. 인베스팅닷컴에서 10 년 만기 미국채 선물 과거 데이터, BTC_USD 비트파이넥스 과거 데이터, CBOE Volatility Index Historical Data, USD_KRW 과거 데이터, WTI 유 선물 과거 데이터, 금 선물 과거 데이터, 천연가스 선물 과거 데이터, 나스닥 종합지수 과거 데이터, 코스피 지수 과거 데이터를 수집한다. 이 보고서에 분석할 대상의 기업은 20 년 4 월부터 23 년 3 월까지 신규 상장한 공모주에 해당한다. 수집한 정보들을 통해서 기업의 이름과 상장일 통해서 하나의 데이터 프레임으로 추가한다. 인베스팅닷컴 데이터를 이용하여 각 종가와 분산에 대하여 5 일, 20 일, 60 일에 대한 이동 평균 컬럼을 추가로 제작한다. 5 일과 20 일 60 일 선정 이유는 5 일은 단기매매선, 20 일은 심리선, 60 일은 수급선으로 실제 주가 분석에서 많이 사용되고 있기 때문이다. 나스닥지수와 코스피 지수 같은 경우는 거래량이 투자 심리와 상당히 밀접해 있다고 생각하여 거래량 컬럼을 추가해 준다. 그리고 수집한 정보들은 종가 기준이기에 상장일 이전에 정보를 반영해야 하므로 하루를 뺀 다음 공모주 데이터프레임과 병합한다. 날짜 컬럼에 대해서는 이전 분석에서 (신규 예탁 지정 종목조회를 통해 상장 수의 계절성 확인) 계절성을 확인하였기 때문에 상장 월에 대한 특성이 있으리라 판단해 월 컬럼을 추가해 준다. 그리고 상장일과 신고서 제출일, 납입일, 상장예정일의 차이를 구하여 새로운 열로 추가한다. 추가한 이유는 계획했던 것과 실제 상장일이 길수록 기업 내부나 외부적 요인이 존재하였고 이는 주가에 영향을 미쳤을 수도 있다고 생각했기 때문이다.

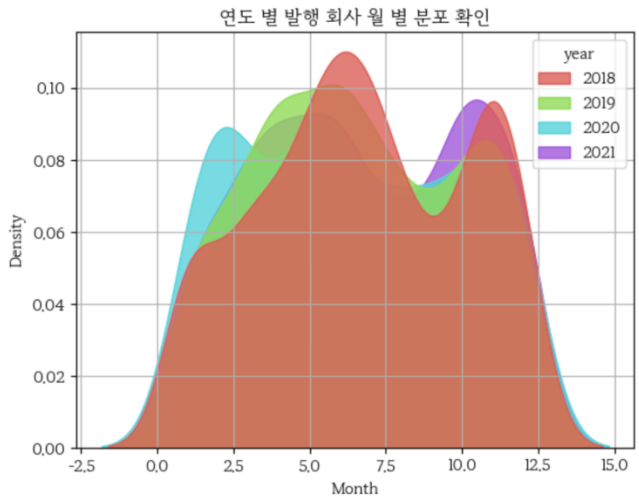

Figure 2 연도 별 월 별 신규 상장 주식 수

문자형 변수인 '시장 구분', '국적', '업종', '증권 구분'의 경우에는 분석을 위하여 더미 변수로 변환하여 진행한다. 증권 구분의 경우는 공모주를 상장할 때 담당 증권이 여러 개가 있는 경우가 많았고, 이를 바로 더미 변수로 하기에는 너무 많은 변수가 생성된다. 그래서 특정 증권이 담당하였는지에 대한 컬럼으로 바꿔서 진행한다. 또한 적절한

분석을 위하여 값이 하나만 있는 경우의 컬럼은 제거한다. 신규상장의 주가에 어떤 것이 중요할지를 생각해 보았을 때 생각한 것은 시장의 흐름이다. 그렇기에 한 기업이 신규 상장했을 경우 이전 신규 상장한 기업들의 상장일 당일 주가가 중요하다고 생각하여 상장일 등락률에 대한 이전 3 개와 이전 7 개가 대하여 이동평균 데이터를 추가한다. 이때 rolling 을 쓸 경우 자신의 값까지 들어가기 때문에 shift 함수를 써서 자신의 종가 데이터가 들어가지 않고 이동평균이 적용되도록 한다. 총 데이터프레임의 모양은 146 개 행의 166 개의 컬럼이다. 여기서 모델에 사용하기 위해 변수들을 표준 정규화를 통해서 정규화를 진행한다. 그리고 성질이 비슷한 컬럼을 분석에 사용하면 공산성의 증가 및 모델의 잘못된 추론 및 모델이 무거워질 수 있기에 변수 간 상관관계 계산을 통하여 0.8 의 강한 상관관계를 나타내면 그 컬럼을 제거하도록 한다. 그렇게 해서 최종적으로 생성된 데이터 프레임의 모양은 146 개의 행과 116 개의 열을 가진다. 컬럼의 종류는 다음과 같다. '종목코드', '공모금액(백만원)', '액면가(원)', '공모가(원)', '매출액', '영업이익' 등을 포함한 146 개이다. 146 개에 대해서는 APPENDIX [1]에 참고한다. 그리고 예측해야 될 라벨 데이터 프레임은 146 개의 행과 13 개의 열을 가진다. 각 컬럼의 통계량 예시는 아래와 같다.

컬럼명 \ 통계값	최댓값	최솟값	중앙값	평균값	1 사분위 수	3 사분위 수
공모금액(백만원)	12,750,000	8,800	29,729	205,858	18,937	53,006
공모가(원)	498,000	1,700	17,000	26,600	11,625	26,375
최초상장주식수(만개)	23,400	244	973	1,773	666	1,510
매출액(백만원)	20,759,117	0	304,452	954,819	130,637	832,725
미국채권 10 년 var	0.84	-1.12	-0.04	-0.07	-0.19	0.14
VIX price	38.02	15.43	21.72	22.75	19.27	25.66
유가 price	115.68	35.79	71.7	69.48	52.21	81.64
나스닥 price	15,993	9,943	12,787	12,803	11,375	13,818
납입일_diff	37	3	6	7.14	6	7
상장일등락률	160	-33	37.6	56.18	2	107.78
6 개월등락률	721.3	-57	10.8	46.99	-7.93	60.15

MODEL

이 보고서에서 사용할 모델은 Logistic Regression, LightGradientBoostingModel, Multi-Layer Perceptron 총 3가지이다. 그리고 Bayesian search기법을 사용한다.

LR(Logistic Regression)은 머신러닝에서 사용하는 분류 알고리즘 중 하나로 주로 이진 분류 문제에 적용되는 모델이다. 입력변수와 가중치의 선형 조합을 시그모이드 함수에 적용하여 예측하며 0.5이상이면 양성 클래스로 0.5이하이면 음성 클래스로 분류하는 모델이다.

LGBM(LightGradientBoostingModel)은 그래디언트 부스팅 알고리즘을 기반으로 한 머신러닝 모델이다. 그래디언트 부스팅은 앙상블 학습의 일종으로 여러 개의 약한 학습기를 결합하여 강력한 예측 모델을 구성하는 방식이다.

MLP(Multi-Layer Perceptron)는 다층 퍼셉트론으로 인공 신경망의 종류이다. MLP는 여러 개의 은닉층으로 구성된 신경망으로, 입력층, 은닉층, 출력층으로 구성되어지며 각 층은 뉴런으로 구성되어 있다. 그리고 뉴런은 가중치와 활성화 함수를 통해 결과를 출력한다. 역전파 학습법을 통해 손실을 최소화시키는 방향으로 학습이 진행된다.

Bayesian search는 머신러닝 하이퍼 파라미터 튜닝에 사용되는 기법으로 확률적인 방법을 사용하여 하이퍼파라미터 공간에서 최적의 조합을 찾는 방식이다.

EXPERIMENTS

위와 같은 데이터프레임으로 진행할 분석은 크게 두가지이다. 1. 등락률에 대한 예측, 2. 매도 시기 결정 예측이며 1.등락률에 대한 예측은 상장일 등락률, 1 개월 등락률, 3 개월 등락률, 6 개월 등락률에 대하여 10%이상 올랐는지, 50%이상 올랐는지에 대한 LR 과 LGBM model 을 사용한다. 이때 LGBM model 의 평가지표는 l1 loss 를 이용하여 진행한다. 그리고 1 개월 등락률, 3 개월 등락률, 6 개월 등락률에 대해서 최신 상장 공모주에 경우 그 정보가 없으니 제하고 진행한다. 모델 학습의 경우에는 train 과 test 를 7:3 으로 나누어 진행한다. Train 으로 학습을 진행한 다음 test 를 이용하여 정확도 및 AUC 를 확인한다. 각 모델의 Accuracy 결과는 다음과 같다.

Accuracy	상장일 등락률	1 개월 등락률	3 개월 등락률	6 개월 등락률
LR 10% 이상	45%	46%	60%	58%
LR 50% 이상	32%	46%	53%	39%
LGBM 10% 이상	64%	56%	50%	47%
LGBM 50% 이상	47%	54%	58%	67%

LGBM model 의 성능이 전반적으로 높았으며 특히 상장일 등락률 10%이상 올랐는가에 대해서 64%의 정확도를 나타내고 있다.

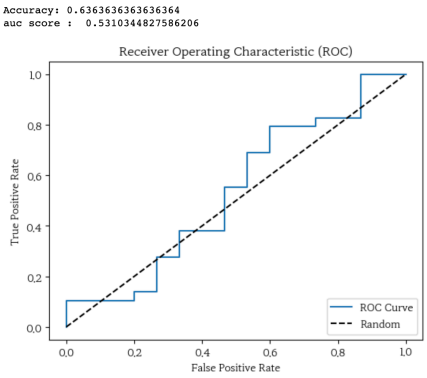

Figure 3 LGBM 10% 이상 ROC-AUC 그래프

두 번째 매도 시기 결정에 대한 분석은 각 주식에 대하여 상장일 등락률, 1 개월 등락률, 3 개월 등락률, 6 개월 등락률 중 가장 높은 것은 라벨값을 설정하며 라벨값은 0~3 까지의 값을 가진다. 그리고 이를 이용하여 LGBM model 과 MLP 딥러닝 모델을 사용하여 4 진 분류를 진행한다. LGBM model 을 사용할 때는 Bayesian search 를 사용하여 3 가지 요소에 대하여 하이퍼 파라미터 튜닝을 추가적으로 진행하였으며 MLP 모델을 사용할 때는 한 개의 은닉층을 사용하여 100 epoch 학습을 진행하였다. 손실 함수는 크로스엔트로피 로스와 옵티마이저는 Adam 을 사용한다. LGBM model 의 경우 53%의 정확도를 나타내며 MLP model 의 경우 47%의 정확도를 나타낸다.

RESULT

위와 같은 모델을 이용하여 공모주 투자를 진행했을 경우, 어떠한 결과가 나타나는지 확인하도록 한다. 모델을 이용함의 유무에 따른, 공모주를 선정하여 판매하였을 경우를 살펴본다. 이 실험은 6 개월 후의 주가를 가지고 있는 데이터를 대상으로 진행하였으며 랜덤적으로 5 개를 선정하여 비교한다.

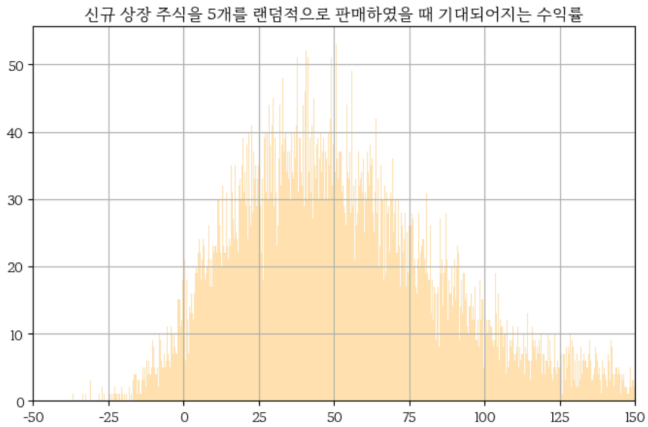

Figure 4 모델 없이 랜덤적으로 투자하였을 경우의 기대 수익률

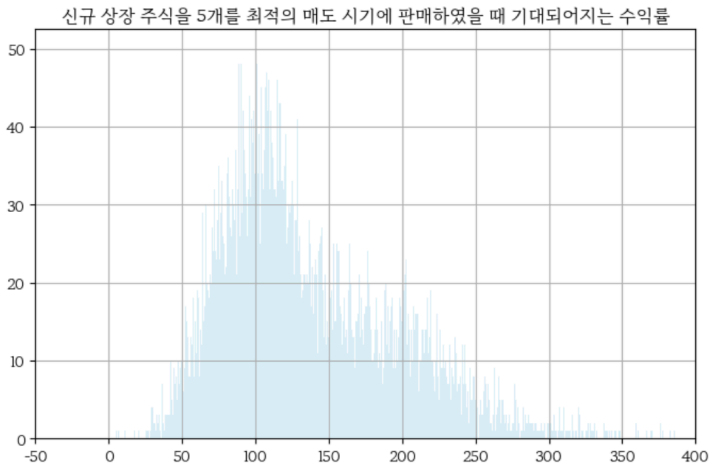

Figure 5 모델을 통한 공모주 선정 및 매도 시기 결정했을 경우의 기대 수익률

Figure 4 은 모델 없이 랜덤적으로 공모주를 선정하고 매도 시기를 랜덤적으로 하였을 경우의 기대 수익률을 나타내는 히스토그램이다. 위의 두 히스토그램 전부다 10000 번을 실행했을 때 평균적으로 나온 5 개 종목에 대한 수익률의 평균값이다. Figure 4 을 평균 내었을 때 값은 58.45%의 값을 보인다. 그리고 Figure 5 는 모델의 결괏값을 통하여 공모주를 선정하고 매도 시기를 결정하였을 때의 기대 수익률을 나타내는 히스토그램이다. . Figure4 를 평균 내었을 때 값은 137.19%의 값을 보인다. 결과를 확인하였을 경우 확실히 모델을 통해 공모주를 선정하고 매도 시기를 결정하였을 경우 기대 수익률이 훨씬 높은 것을 볼 수 있다. 다만, 위의 자료 둘 다 높은 기대 수익률을 가지고 있는데 이는 21~22 년도 주식 특성상 코로나가 종료되고 나서 주식 시장의 회복 및 상승세 그리고 공모주 열풍이 불어 전반적으로 높은 값을 보인다. 또한, 데이터 수급의 부족으로 인하여 학습하는 데 사용한 데이터를 결괏값 추론에 사용하였기에 모델의 과적합성이 존재할 수 있다.

CONTRIBUTION

우리는 지금까지 공모주의 불패 신화의 진위 그리고 공모주 선정 및 공모주 매도 시기 결정에 대해 분석했다. 공모주의 불패 신화는 코로나 시기 20~21 년도 공모주의 주가 흐름을 봤다시피 모든 공모주가 우리에게 이득을 가져다주는 것은 아니며 내부 및 외부 환경을 고려하며 투자해야 함을 알 수 있다. 또한, 공모주 선정 및 매도 시기 결정에 있어 여러가지 요소들을 고려하여 분석한다면 보다 유리한 포트폴리오를 구성할 수 있음을 있다. 이를 통해 새롭게 미래의 공모하는 주식이 생긴다면 이와 같은 과정을 통해서 적합한 공모주인지를 분석하고, 매도 시기를 결정하는 데 도움을 줄 수 있을 것이다.

APPENDIX

[1] (컬럼 이름 리스트 146 개) '종목코드', '공모금액백만원', '액면가원', '공모가원', '매출액', '영업이익', '미국채권 10 년 20MA_line', '미국채권 10 년 60MA_line', '비트코인 price', '비트코인 var', '비트코인 5MA_line', '비트코인 20MA_line','비트코인변동 5MA', 'VIXprice', 'VIXvar', 'VIX5MA_line', 'VIX20MA_line', '환율 var', '환율 5MA_line', '환율 20MA_line', '환율 60MA_line', '유가 price', '유가 var', '유가 5MA_line', '유가 20MA_line', '금 price', '금 var', '금 60MA', '금 5MA_line', '금 20MA_line', '금 60MA_line', '금변동 5MA', '천연가스 var', '천연가스 5MA_line', '천연가스 20MA_line', '천연가스 60MA_line', '천연가스변동 5MA', '나스닥지수 var', '나스닥지수 5MA_line', '나스닥지수 20MA_line', '나스닥지수 60MA_line', '나스닥지수거래량', '코스피지수 var', '코스피지수 5MA_line', '코스피지수 20MA_line', '코스피지수 60MA_line', '코스피지수거래량', '코스피지수변동 5MA', '신고서제출일_M', '수요예측일정_M', '신고서제출일_diff', '납입일_diff', '주관사_갯수', '삼성증권_isin', '현대차증권주식회사_isin', '골드만삭스증권서울지점_isin', '미래에셋증권주식회사_isin', '유안타증권_isin', 'KB 증권_isin', '신영증권_isin', '한화투자증권_isin', '키움증권_isin', 'DB 금융투자주식회사_isin', '엔에이치투자증권주식회사_isin', '제이피모간증권회사서울지점_isin', '신한투자증권주식회사_isin', '한국투자증권_isin', 'IBK 투자증권_isin', '하나증권주식회사_isin', '대신증권_isin', '시장구분_유가증권상장추진', '국적_대한민국', '국적_홍콩', '업종_1 차', '업종_가구', '업종_건물', '업종_건물설비', '업종_건축기술', '업종_광고업', '업종_구조용', '업종_그외', '업종_금융', '업종_기초', '업종_기타', '업종_반도체', '업종_상품', '업종_섬유', '업종_소프트웨어', '업종_알코올음료','업종_영화', '업종_오디오물', '업종_운동', '업종_운송장비', '업종_음식료품', '업종_의료용', '업종_의료용품', '업종_의약품', '업종_일반', '업종_일차전지', '업종_자동차', '업종_자연과학', '업종_전기', '업종_전동기', '업종_전자부품', '업종_측정', '업종_컴퓨터', '업종_통신', '업종_특수', '업종_플라스틱제품', '상장일등락_이동평균_3', '상장일등락_이동평균_7'

Datasets 출처

출처	내용	링크
대한민국 대표 기업공시채널 KIND	공모기업현황, 신규상장기업현황, 영업성과추이, 일반재무현황, 종목별공모가대비주가등락률현황	https://kind.krx.co.kr/main.do?method=loadInitPage&scrnmode=1
인베스팅 닷컴	10년만기 미국채 선물 과거 데이터, BTC_USD 비트파이넥스 과거 데이터, CBOE Volatility Index Historical Data, USD_KRW 과거 데이터, WTI유 선물 과거 데이터, 금 선물 과거 데이터, 천연가스 선물 과거 데이터, 나스닥종합지수 과거 데이터, 코스피지수 과거 데이터	https://www.investing.com
공공데이터 포털	금융위원회_주식시세정보, 신규예탁지정종목조회	https://www.data.go.kr/iim/api/selectAPIAcountView.do