랜덤포레스트를 활용한
비상장기업의 적정 주가 예측가능성 연구

팀명 : 상장하장

팀원 : 신재욱, 김현호

논문초록

본 연구에서는 비상장기업의 적정주가를 기업의 재무데이터를 비롯하여 사업보고서에 기초한 자료로 예측할 수 있을지 알아보았다. 특히 과적합 문제가 거의 발생하지 않고 높은 정도를 가진 것으로 알려진 랜덤포레스트를 활용하였다.

예측 결과, Test set 으로부터 측정한 RMSE 는 약 13,691 원을 기록하였다. 이는 주가의 분포가 304 만원부터 94 원까지 분포되어 있는 것을 감안할 때 상당한 정확도를 보이는 것으로 판단할 수 있다. 또한 주가가 일정한 값 이하일 때만을 대상으로 학습과 예측을 할 경우 이러한 정확도는 크게 향상되었으며 특히 3 만원 이하 기업에 대한 예측에서 RMSE 는 3,305 원을 기록하였다.

이러한 모형이 비상장기업으로 일반화시킬 수 있을지에 대해 확인하기 위해 2017 년 말부터 2018 년 초 사이에 상장한 기업을 대상으로 예측을 실시하여 공모가와 정확도를 비교하였다.

비교결과 모형 예측 RMSE 는 약 6,346 원을 기록하였고, 공모가 RMSE 는 약 4,691 원을 기록하였다. 공모가가 약간의 더 높은 정확도를 기록하였으나, 공모가 선정에 있어서 많은 시간과 비용이 투입된다는 것을 감안할 때, 본 연구 모형이 실제 업무에서도 활용될 수 있는 여지가 높다고 할 수 있다.

목차

제 1장 서론

제 1절 연구배경

주식은 크게 상장주식과 비상장주식으로 구분할 수 있다. 상속세 및 증여세법(이하 '상증법')과 상속세 및 증여세법 시행령에 따르면 유가증권 시장과 코스닥시장에 상장된 주식을 '상장주식'으로 분류하고, 이외에 장외시장에서 거래되는 주식을 비상장주식이라 정의하고 있다(박근우 외, 2018).

상장주식은 장내시장에서의 거래를 통해 시세가 정해지는 반면, 비상장주식의 경우 장외시장에서 거래되므로 거래량의 부족과 신뢰 문제 등 다양한 이유로 인해 객관적인 가치를 산정하기 힘들게 된다. '상증법'에서 재산의 평가는 시가평가를 원칙으로 하고 있지만, 현실적으로 시가 산정이 어려운 경우가 많다. 이 경우 당해 재산의 종류·규모·거래상황 등을 감안하여 재산별로 규정된 보충적인 방법(이하 '보충적평가법')에 의하여 평가하도록 하고 있는데, 이러한 보충적평가법은 논리가 희박하고, 평가액이 실제 주식가치를 반영하지 못한다는 주장이 제기되어 왔다(김병수·오현탁, 2011). 또한, 보충적평가법은 기업의 규모 업종 및 기업의 경영방침, 자산보유규모, 수익창출정도를 고려하지 않고 모든 기업에 무차별적으로 적용되기 때문에 평가가치가 실질가치와 큰 차이를 보일 가능성이 매우 높다. 이 때문에 실제 납세자와 과세당국사이에 끊임없는 분쟁을 유발하였다(이병철 외, 2005).

비상장주식의 가치평가는 일반적으로 ①비상장기업이 신규로 상장하는 경우, ②비상장기업이 전환사채 또는 신주인수권부사채를 발행하는 경우, ③상장기업과 비상장기업이 합병하는 경우, ④상장기업과 비상장기업이 주식을 교환하는 경우, ⑤상장기업이 보유한 비상장기업의 주식을 공정가치로 평가하는 경우 등의 다양한 상황에서 이루어진다(박종찬·최원석, 2011).

이 때문에 적정 주가를 정확히 평가하는 것은 해당 회사와 주식거래자뿐만 아니라, 증권사, 은행, 보험사, 회계사 등 다양한 주체들에게 중요한 일이라 할 수 있다.

그러나 그동안 머신러닝을 활용한 많은 연구가 주식시장의 지수 자체를 예측하거나, 상장주식의 등락을 예측하기 위해 이루어져 왔을 뿐, 비상장주식에 대한 평가를 위해서는 연구가 미진한 상황이다. 따라서 머신러닝 기법을 응용하여 비상장기업의 적정 주가를 평가하는 것은 학술적 측면과 실용적인 측면 모두에서 기여도가 있을 것이다.

효율적 시장가설(Efficient Market Hypothesis, EMH)에 따르면 현재 공개된 정보 중에서 주가에 영향을 줄 수 있는 정보는 정확하고 신속하게 주가에 반영된다. 따라서 비정형 정보가 주가에 미치는 영향이 평균적으로 상쇄된다고 할 때, 상장 기업의 주가는 기업의 재무 정보를 반영한 내재가치에 가까울 것이고, 이를 학습해 비상장기업의 재무 정보를 활용하는 것은 높은 정확도를 갖고 적정 주가에 근접한 값을 예측해낼 수 있을 것으로 기대된다.

또한, 본 연구에서는 상장 기업의 재무자료에 머신러닝 기법 중 하나인 랜덤포레스트(Random Forest)를 적용하여 학습하고 이를 통해 비상장기업의 적정 주가를 평가하고자 한다. 랜덤포레스트는 타 머신러닝 기법과 비교해 높은 정확도와 안정성을 갖추었으며, 과적합 문제가 거의 발생하지 않고 변수 소거에 있어서 연구자 개인의 주관이 개입될 여지가 적은 등 많은 장점이 존재함에도 불구하고, 그간 주가와 관련한 예측 연구에 있어서 Eo et al(2017)를 비롯한 몇몇 연구를 제외하고는 활발히 활용되지 않았다. 또한 이러한 연구들도 분류기(Classifier)로서 랜덤포레스트를 활용하고 있을 뿐, 리그레서(regressor)로서 주가를 예측(Prediction)하는데 활용되지 않았다.

따라서 본 연구에서는 랜덤포레스트를 활용하여 기업 재무데이터와 거시데이터를 학습하고, 이를 통해 구한 비상장기업의 적정주가에 대한 예측값이

증권사가 평가한 공모가와 비교해서 적정 주가에 얼마나 근접하게 예측하고 있는지에 대해 알아보고자 한다.

제 2절 선행연구

본 연구와 관련하여 선행 연구는 크게 머신러닝을 통한 주가 예측을 위한 연구와 비상장기업의 주가를 평가하기 위한 연구로 나눌 수 있다.

먼저 머신러닝을 통한 주가 예측을 위한 시도로 이재원(2013)은 Multi Layer Perceptron을 사용하여 14%이상의 주가 상승을 고변동 주가 패턴으로 정의하고 일본식 봉 차트와 이동평균선을 적용하였다. 해당 연구에서는 각각의 패턴별로 독립된 신경망을 사용하여 정확도를 높이고자 하였다.

Ashwin Siripurapu(2015)는 데이터의 특징을 추출하여 가중치의 개수를 줄여 계산속도가 빠르다는 장점을 지닌 Convolution Neural Network 알고리즘을 사용하였다. 해당 연구에서는 주가 정보를 나타내는 날짜, 거래량, 시작가, 종가, 상한가, 하한가등의 시계열데이터를 2차원 그래프로 변환하여 주가를 예측하였다.

양진용(2016)은 머신러닝 방법으로 재무정보를 활용해 주식 가격의 예측력을 검증하였다. 이를 통해 회사의 내재 가치를 나타내는 재무정보가 주식 가격 예측에 얼마나 효과적인지 평가하고자 하였다. SVM과 인공신경망, 결정나무, 적응형부스팅을 통한 예측 성능을 비교하였으며, SVM의 성능이 가장 우수한 것으로 나타났다. 또한, 전문가 예측을 점수화하여 이를 SVM의 예측 결과와 비교하였으며, 그 결과 SVM의 주가 예측력이 전문가 예측에 비하여 우수하게 나타났음을 확인하였다. 한편, 재무정보를 활용한 예측이 단기간에는 우수하지만 기간이 길어질수록 예측력이 떨어지는 것으로 나타났다고 지적하였다.

비상장기업의 주식의 가치평가를 위한 연구에서 박종찬 외(2011)는 삼성

에버랜드의 전환사채 발행사례를 활용하여 여러 가치평가모형의 실제 적용 방법을 알아보고, 가치평가모형별 장단점과 논란의 소지가 될 수 있는 요인들을 분석하였다. 구체적으로 유사기업이용법, 과거거래이용법, 배당할인모형, 현금할인모형, 초과이익모형, EVA모형, 자산접근법에 대해서 분석하였으며, 이를 토대로 가치평가방법에 대한 올바른 이해와 한계점에 대하여 논하였다.

한편 김병수 외(2011)는 상속세법상 비상장주식의 보충적평가법이 실제주가와 괴리되어 있고, 대안으로 제시된 유사기업비교평가법이 지나친 제약조건으로 실제 평가에서 거의 사용되지 못한다는 것에 문제를 제기하였다. 이에 시장지배력모형 일부를 수정한 모형을 제시하고, 기존 모형과 대안모형의 주가예측력을 실증분석을 통해 비교 평가하였다. 이를 통해 유사매출모형과 결합매출모형이 다른 모형들에 비해 우수한 주가 예측력을 보여준다는 것을 확인하였다.

제 2장 본론

제 1절 랜덤포레스트

지금까지 많은 연구에서 주가 예측을 위한 머신러닝 방법으로서 주로 로지스틱 회귀분석(LOGIT)이나 퍼셉트론(Perceptron), 의사결정나무(Decision Tree), 인공신경망(Artificial Neural Network, ANN), SVM(Support vector Machine) 등이 주로 활용되어왔다. 그러나 이러한 기법들은 종종 독립변수를 선정하는데 있어서 해답을 주지 못하고, 이상치(Outlier)에 취약하거나 과적합(Overfitting) 문제가 발생할 수 있다. 또한 많은 모수(tuning parameter)에 대한 조정 작업이 요구된다는 문제를 지적받아왔다(안현철 외, 2016).

이에 본 연구에서는 비상장기업의 적정 주가 예측을 실시하기 위해 머신러닝 기법 중 하나인 랜덤포레스트(Random Forest)를 적용하고자 한다. 랜덤포레스트 알고리즘은 비교적 '약한' 학습기인 트리(Tree)를 기반으로한 앙상블(ensamble)을 만드는 방법으로 Breiman(2001)에 의해 처음 개발되었다. 이처럼 트리를 기반으로 앙상블을 만듦으로써, 개별 트리의 정밀도는 떨어질 수 있으나, 이를 종합하여 예측을 수행하는 랜덤포레스트의 정확도와 안정성을 높일 수 있다.

특히 랜덤포레스트는 대수의 법칙에 의해 숲의 크기(개별 트리의 수)가 커질수록 일반화 오류가 특정 값으로 수렴하게 되어 과적합화를 피할 수 있으며, 각 개별 트리들을 학습시킬 때 전체 학습용 자료에서 무작위로 복원 추출된 데이터를 사용하고 있어 잡음이나 이상치로부터 크게 영향을 받지 않는 것으로 알려져있다(한은정, 2004). 이러한 배경에서 랜덤포레스트가 비상장기업의 적정 주가 예측에 가장 합리적이라 판단되어 이를 활용하고자 한다.

제 2절 모형 및 변수 소개

예측을 위해 2011년 1분기부터 2017년 4분기까지 우리나라 유가증권시장과 코스닥시장에 상장한 기업 중 금융업이 아닌 업종에 종사하는 기업을 대상으로 재무 데이터를 사용하였다(구체적인 대상 업종은 <부록1> 참고). 비상장기업은 장내시장에서 거래되지 않는 상태이지만, 본 연구에서는 이들 기업이 장내시장에 상장할 경우를 가정할 때의 적정주가를 예측하고자 하기 때문이고, 이렇게 함으로써 증권사가 평가한 공모가와의 비교를 통해 적정주가의 예측의 실용성을 가늠해볼 수 있을 것이기 때문이다.

재무 데이터는 한국상장회사협의회에서 제공하는 KOCOinf(Korea Company Information)에서 추출하였으며 구체적으로 재무상태표를 비롯하여 손익계산서, 현금흐름표, 자본변동표에 나타나있는 계정 항목들을 중복한 경우를 제외하고 최대한 사용하고자 하였다. 또한 사업보고서에 나타난 일반사항 중 산업명, 설립일, 종업원 수, 발행주식수를 변수로 추가하였다. 그 결과 KOCOinf을 통해 약 30개의 설명변수를 추출할 수 있었다.

이 중 산업명은 해당 기업이 속해 있는 산업을 나타내고 있다. 따라서 이를 더미변수로 바꾸어 처리해주었으며 구체적인 항목은 <부록 1>에 추가하였다. 이와 함께 1~4분기에 해당하는 더미변수도 추가로 통제하였다.

마지막으로 시계열 추세를 통제해주기 위해 거시 변수들을 추가로 통제하였다. 데이터는 한국은행 경제통계시스템을 통해 구하였으며, 구체적으로 분기별 경제성장률, 총저축률, 제조업평균가동률, 실업률, 코스피지수, 경상수지 등이 있다.

그 결과, 기업 재무데이터와 거시 데이터를 모두 합하여 총 88개의 독립변수를 입력변수를 설정하였다. 구체적인 변수명은 다음과 같다.

모형 독립변수
설립일, 종업원, 당좌자산, 재고자산, 비유동자산, 유동부채, 비유동부채, 자본금, 자본잉여금, 자본조정, 기타포괄손익누계액, 이익잉여금, 매출액(영업수익), 매출원가, 매출총이익(손실), 판매비와관리비(영업비용), 영업이익(손실), 영업외수익, 영업외비용, 법인세비용차감전(계속사업)손익, (계속사업손익)법인세비용, 계속사업이익(손실), 당기순이익(순손실), 영업활동으로 인한 현금흐름, 투자활동으로 인한 현금흐름, 재무활동으로 인한 현금흐름, 환율변동으로 인한 차이조정, 현금의 증가(감소), 경제성장률, 총저축률, 제조업 평균가동률, 실업률, 코스피지수, 경상수지, 발행주식수, 산업별 더미변수 51개 등

재무데이터를 이와 같이 사용할 경우, 몇몇 변수들의 정의상 반드시 상당히 높은 공선성을 갖게 된다. 이러한 문제를 해결하기 위해 데이터 학습에 앞서서 재무자료에 대해서 주성분분석(Principal Component Analysis, PCA)을 통한 차원축소를 진행하였다. PCA를 통해 재무자료등 기업의 미시자료의 90% 이상을 설명하는 주성분 3개를 선택하였고, 거시자료의 90%이상을 설명하는 주성분 1개를 선택하였다.

한편 우리가 예측하기를 원하는 것은 기업의 적정 주가이므로 종속변수는 각 기업의 주가를 의미한다. 특히 본 연구에서는 기업의 사업보고서가 나오는 해당 월의 주가를 산술평균한 값을 사용하였다. 특정한 날의 종가를 사용하지 않고 월별 평균자료를 사용함으로써 평균적인 주가 예측에 있어서 정확도를 향상시킬 수 있을 것으로 기대된다. 그 결과 기본 모형은 (1)과 같다.

(1) $주가_{i,t} = \beta_0 + x_{i,t} + \gamma_t + \beta_k D_{k,i} + u_{i,t}$

여기서 $x_{i,t}$는 기업의 사업보고서를 기반으로 한 재무 변수와 그 외에 기업과 관련한 변수들을 의미한다. γ_t는 거시데이터를 나타내며, $D_{k,i}$는 $k+1$개의 산업명을 더미변수로 나타낸 것이다.

한편, 모형(1)에서 차원축소를 통해 최종적으로 사용될 모형은 다음과 같다.

(2) $\text{주가}_{i,t} = \alpha_0 + \alpha_1 P1_{i,t} + \alpha_2 P2_{i,t} + \alpha_3 P3_{i,t} + \delta P4_t + \varepsilon_{i,t}$

여기서 $P1, P2, P3$는 모형(1)에서 $x_{i,t}$에 해당하는 변수들을 차원축소하여 구한 주성분이고, $P4$는 모형(1)에서 거시변수들을 의미한 γ_t를 차원축소하여 구한 주성분을 의미한다.

유가증권시장과 코스닥 시장에 상장한 1,978개 회사에 대하여 2011년 1분기부터 2017년 4분기까지 28분기에 해당하는 자료를 구축하였으며, 일부 결측치를 제외하고 총 44,808개의 관측치를 확보하였다.

제 3절 기초통계량

주요 기초통계량은 다음 <표1>과 같다.

<표1> 종속변수(주가) 기초통계량

	전체	대기업	중견기업	중소기업
관측치	45,482	5,774	20,227	19,416
평균	27,655.4	102,957.4	24,371.7	8,766.5
표준편차	103,881.9	244,959.3	69,131.5	20,051.9
최솟값	94	198	133	94
1분위수	2,906	8,028	3,380	2,300
중위값	6,296.5	30,133	7,558	4,358.5
3분위수	16,747.75	86,332.5	18,956	8,742
최대값	3,039,182	3,039,182	1,196,190	604000

<표1>에서 종속변수인 주가의 분포를 보면 최대 303만원부터 최소 94원까

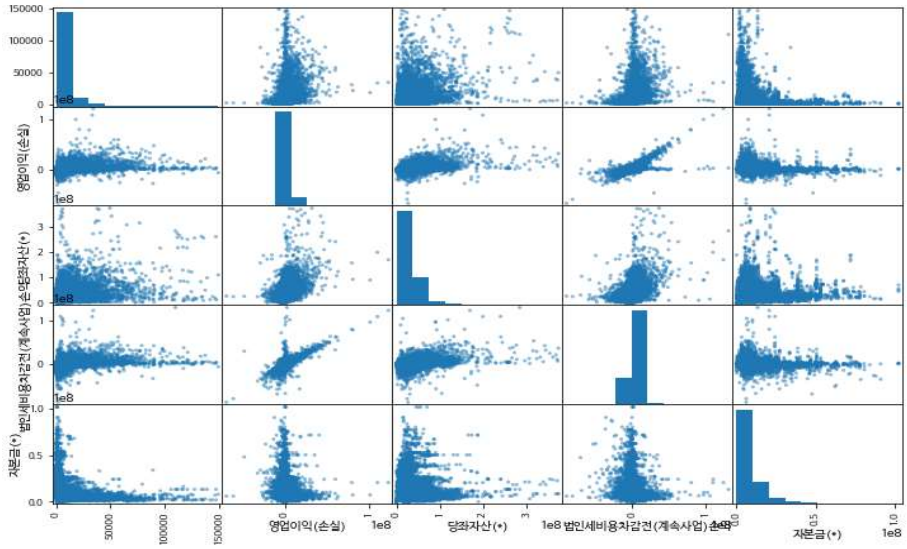
지 굉장히 큰 범위를 보이고 있다. 그러나 중위값은 6천원 수준이고 3분위에서도 2만원이 채 되지 않을 정도로 데이터 편중이 심하다는 것을 알 수 있다. 이에 따라 정규화를 위해 종속변수에 로그를 취하여 예측을 진행하였다.

<표2> 종속변수와의 상관관계

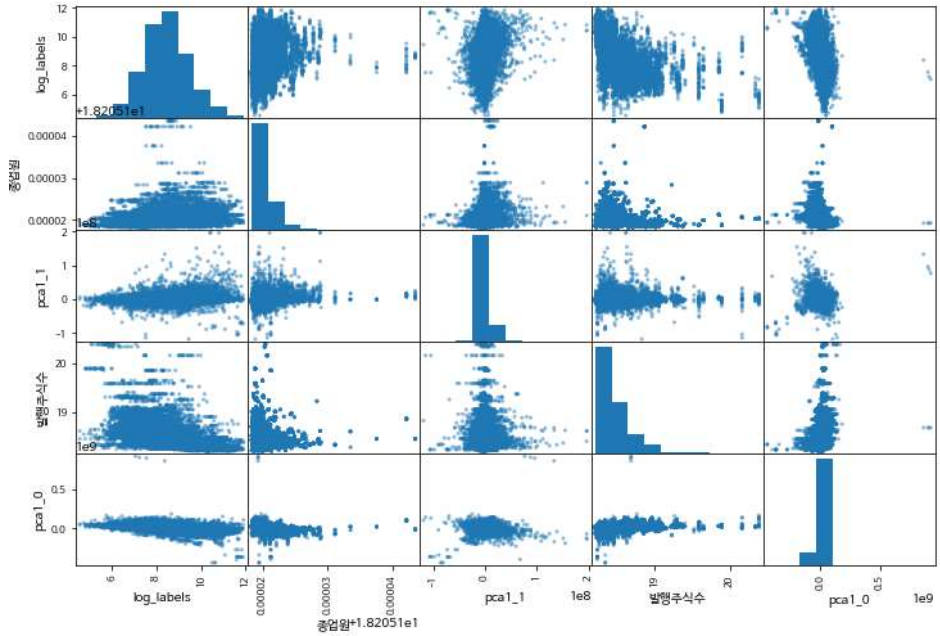
당좌자산	0.3142	판관비	0.1346	재고자산	0.0383
영업이익	0.2859	총저축률	0.1121	유동부채	0.0068
이익잉여금	0.2699	매출액	0.1031	환율변동 차이조정	-0.0103
법인세비용 차감전손익	0.2675	경상수지	0.0791	경제성장률	-0.0126
당기순이익	0.2552	현금의증감	0.0787	영업외비용	-0.0222
계속사업 이익	0.2544	재무활동 현금흐름	0.0723	자본조정	-0.1041
매출총이익	0.2379	코스피지수	0.0610	제조업 평균가동률	-0.1187
법인세비용	0.2089	영업외수익	0.0602	투자활동 현금흐름	-0.1463
자본잉여금	0.1827	비유동부채	0.0563	발행주식수	-0.1616
영업활동 현금흐름	0.1771	기타포괄 손익누계액	0.0535	자본금	-0.1978
비유동자산	0.1543	실업률	0.0531		
종업원	0.1537	매출원가	0.0443		

위 <표2>를 살펴보면 당좌자산이 종속변수인 주가와 가장 높은 상관관계를 갖고 있음을 확인할 수 있었다. 그리고 이익관련 자료와 자본잉여금 등이 뒤를 이었다. 반면 자본금의 경우 주가와 음(-)의 상관 관계를 갖고 있다는 것을 확인할 수 있다. 주요 변수에 대해 시각화한 결과는 다음 <그림1>과 같다.

<그림1>종속변수와 주요 독립변수와의 상관도



<그림2> 종속변수 로그값과 모델 PCA 후 변수와의 상관도



마지막으로 <그림2>는 종속변수를 로그변환하고 PCA를 적용한 상태에서의 상관도를 나타내고 있다.

제 4절 예측 결과 및 평가

학습된 모형의 정확도를 평가하기 위해 전체 데이터의 80%를 Train set으로 분류하였고 나머지 20%를 Test set으로 분류하여 예측의 정확도를 평가하였다. 또한 랜덤포레스트의 경우 몇 가지 Tuning parameter들을 갖고 있으므로, 10-folds Cross validation을 통해 최적 모수를 설정하였다. 구체적으로 n-estimator의 경우 55이고, Max depth는 40이다. 평가의 기준이 되는 값은 MSE(Mean Squared Error)의 제곱근을 사용하였다. 랜덤포레스트를 활용한 예측 결과는 다음과 같다.

<표3> 예측결과 RMSE

	RMSE
Train set	16원
Test set	38,908원

<표3>을 보면 In sample에서의 오차를 의미하는 Train RMSE는 16원으로 나타난 반면, 학습에 사용하지 않은 Test set에서의 RMSE는 38,908원으로 나타났다. 이는 랜덤포레스트의 장점과 차원축소, 정규화 등 여러 가지 전처리과정을 거쳤지만, 여전히 과적합 문제가 나타나고 있음을 보여주고 있다. 그러나 <표1>에서 확인하였듯이 종속변수인 주가의 범위가 3,039,182부터 94원까지 넓게 주어져 표준편차가 103,882원인 것을 고려하면 이러한 결과를 비관적으로만 생각하기는 어렵다.

이에 추가적으로 데이터를 중소기업으로 한정하고, 추가로 일정값 이하의 주가를 가지도록 제한을 설정하여 예측과 평가의 과정을 다시 수행하였다. 이것은 좀 더 실용적인 측면에서 지지를 받을 수 있을 것으로 기대된다. 왜냐하면, 머신러닝 방법으로 적정주가를 예측하고자 함은 시가가 존재하지

않는 경우에 활용하고자 하는 목표가 크고, 시가를 확인하기 어려운 비상장 기업은 대부분 중소기업이면서 낮은 적정 주가를 형성하고 있을 가능성이 높기 때문이다. 따라서 약간의 제약을 통해 더 높은 정확도를 얻을 수 있다면, 실제 업무에서 이를 활용할 수 있을 가능성도 높아질 것이다.

이를 정리한 결과는 <표4>에 나타난 것과 같다.

<표4> 중소기업 대상 예측결과

	RMSE
기본모형	13,691원
15만원 이하	6,6626원
5만원 이하	5,003원
4만원 이하	4,138원
3만원 이하	3,305원

*중소기업으로 분류된 기업만을 대상으로 데이터를 학습해 구한 결과이다.
1번째 열은 종속변수에 대한 제약을 의미하는 것으로 3만원 이하는
중소기업 중 주가가 3만원 이하인 기업을 대상으로 학습을 했다는 것을 의미한다.
RMSE는 Test set에 대한 MSE의 제곱근 값이다.

<표4>를 보면 Test set에 대하여 구한 RMSE가 <표3>과 비교해서 대폭 낮아졌다는 것을 알 수 있다. 이는 중소기업만을 한정해 학습하여 주가를 예측할 경우 예측의 성능을 훨씬 더 높일 수 있음을 보여준다. 특히 <표1>에서 살펴본 것처럼 최대값이 60만원이 넘고, 최소값이 94원이고 표준편차가 20,051원인 상황에서 13,691원의 RMSE는 랜덤포레스트의 예측 성능이 우수함을 보여준다.

추가적으로, 주가가 일정 수준보다 낮은 경우만을 추출한 경우 예측의 정확도가 개선되어 3만원 이하의 주가를 가진 기업만을 대상으로 예측을 실시하였을 때, RMSE가 3,305원으로 나타났다. 3만원 이하로 제약을 설정하는 것은 신규상장 기업의 공모가액이 대부분 3만원 이하에서 형성한다는 것을

고려할 때, 그다지 강한 조건은 아니라고 볼 수 있다.

제 5절 비상장기업에 대한 예측 가능성

본 절에서는 시행한 예측 모델이 비상장기업에 대해서도 사용가능한지에 대해 평가하기 위해 신규상장기업에 대하여 상장 시 예측값을 구해 그 성능을 평가해보고자 한다.

특히 이것이 얼마나 실용적인 결과인지를 알아보기 위해 신규상장기업이 상장 시 주관 증권사가 평가한 적정 주식의 가치인 '공모가'와 비교하고자 한다. 따라서 먼저 기업의 상장 과정에 대해 소개하고, 전문가 평가 정확도를 RMSE로 산출한 결과를 보여줄 것이다.

일반적으로 비상장기업이 상장할 때는 기업공개를 통해 일련의 절차를 밟아야 한다. 이 과정에서 공개될 기업의 가치를 시장가치에 근접하게 평가하여 투자자에게 제시해야 하는데 이 때문에 기업공개 절차를 전문적으로 수행할 증권회사를 정해 기업공개 주관업무 계약을 맺게 된다. 증권사는 이후 기업 실사를 진행하고, 기업 실사 결과를 기초로 사업보고서와 유가증권신고서를 검토하게 된다. 이후 이를 금융당국에 제출하여 회사의 사업성에 대해 투자자에게 알리게 된다.

이후 증권회사는 수요예측 과정(Bookbuilding process)을 통해 투자자의 수요정보를 수집한다. 즉 기관투자자를 대상으로 이들이 희망하는 주식의 가격과 수량을 접수하고, 이를 기초로 기업과 협의한 공모가를 최종결정하게 된다. 이 때문에 수요예측 방식의 공모가 결정에는 증권회사와 기관투자가의 역할이 중요해진다. 이와 같은 공모가 결정방식은 우리나라를 비롯해 미국, 유럽 등 대부분의 나라에서 채택하고 있는 방식이다.

공모가가 시장가격보다 낮게 설정될 경우 공모에 참여한 투자자에게 이득이 되며, 주주들의 이익에는 반하게 된다. 그러나 일반적으로 주관 증권사

의 수수료는 공모가에 비례하여 책정되기 때문에 증권사에게 공모가를 시장 평가액보다 낮게 설정할 유인이 적다. 또한 공모가가 시장가격보다 높게 설정될 경우 공모에 참여한 투자자들에게 손해를 입히게 되므로 해당 증권사는 차후에 주관업무를 맡기 어려워질 수 있어 시장 평가액보다 공모가를 높게 설정할 유인도 적다고 할 수 있다.

공모가가 얼마나 정확하게 이루어졌는지를 보고 우리의 예측 모델과 비교하기 위해 공모가의 정확도를 평가하고자 한다. 이를 위해 2017년 연말부터 2018년 초에 상장한 기업들을 뽑아 '공모가'와 '상장 1개월 후 주가'에 대해 RMSE를 도출하였다. 여기에서 '상장 1개월 후 주가'를 적정 주가의 대리변수로 사용하였는데, 이는 더 짧은 단기에는 시장에서 정확한 평가가 이뤄지지 않을 수 있고, 시간이 너무 지나게 되면 상장 당시의 상황이 아닌 추가적인 기업 내·외부의 상황으로 주가가 형성될 가능성이 높기 때문이다. 또한 연구자의 임의성을 최소화한다는 측면도 상징적인 1개월이라는 기준을 적용하고자 하는 이유이다.

공모가의 정확도를 평가하기 위해 선정한 기업공시채널인 KIND에서 대상기업을 선정하였으며 구체적인 이름은 다음과 같다.

공모가 정확도 평가 대상 기업
디바이스이엔지, 배럴, 비즈니스온커뮤니케이션, 세종메디칼, 시스웍, 에스엔피월드, 에코마이스터, 엠플러스, 유티아이, 제노레이, 덕우전자, 신흥에스이씨, 씨앤지하이테크, 알리코제약, 에스트래픽, 영화테크

공모가의 정확도와 우리 모형의 예측값을 나타내는 RMSE는 다음과 같다. 구체적인 수식과 결과는 다음과 같다.

<표5>연구모형과 공모가 정확도 비교

산술식	$\sqrt{(\text{Mean}[(\text{적정주가} - \text{예측값})^2])}$	$\sqrt{(\text{Mean}[(\text{적정주가} - \text{공모가})^2])}$
결과	약 6,346원	약 4,691원

<표5>의 결과를 살펴보면 먼저 우리 모형을 통해 2017년말~2018년초 상장한 기업의 주가를 예측했을 때, 적정주가와의 평균적인 차이를 의미하는 '예측값 RMSE'가 약 6,346원으로 나타났다. 반면 공모가와 적정주가의 평균적인 차이를 의미하는 '공모가 RMSE'는 약 4,691원으로 나타났다. 이는 동일 기업을 대상으로 했을 때, 공모가가 적정주가를 좀 더 잘 반영하고 있음을 의미한다. 이것은 공모가가 사업계획서를 기반으로 기업 실사를 거쳐 실질적인 수요예측을 실시하기 때문으로 보인다.

그러나 이러한 공모가 평가 작업이 많이 시간과 비용을 투입했을 때의 결과라는 점을 주의할 필요가 있다. 기업 사업보고서를 기반으로 머신러닝으로 예측한 값과 공모가의 정확도가 큰 차이가 없다면, 이러한 결과는 빠른 시간과 낮은 비용으로 비상장기업의 주가를 평가해야할 때, 머신러닝이 사용될 수 있다는 의미를 내포한다고 볼 수 있다.

제 3장 결론

본 연구에서는 랜덤포레스트를 활용하여 유가증권시장과 코스닥시장에 상장한 기업의 사업보고서에 기반한 데이터와 거시데이터를 활용하여 비상장기업의 적정 주가 예측가능성에 관하여 알아보았다.

예측 결과 랜덤포레스트를 활용한 기업의 주가 예측은 총 편차와 비교해서 높은 수준의 정확도를 갖고 있다는 것을 확인할 수 있었다. 이러한 결과는 특히 주가가 특정 값 이하인 기업을 대상으로 하거나 중소기업만을 대상으로 했을 때 두드러졌다. 이러한 결과가 비상장 기업의 주가 예측에도 사용될 수 있을지를 가늠하기 위해 최근 상장한 기업을 대상으로 증권사가 정한 공모가와 본 연구 모형의 예측값들이 상장 1개월 후 주가를 얼마나 정확히 맞추는지를 평가하였다. 그 결과 우리 모형의 예측 RMSE는 약 6,346원를, 공모가 RMSE는 약 4691원을 기록하였다.

이러한 결과는 다음 학술적 측면과 실용적 측면에서 다음과 같은 가치가 있다.

첫째, 그동안 주된 연구가 증권시장의 주가지수의 등락을 예측한 데 반해 본 연구는 구체적인 주식 가격을 산출하였고, 머신러닝이 특정 값을 예측하는데 있어서도 전문가 평가에 준하는 높은 성능을 보일 수 있다는 것을 확인하여 학술적 의의가 있다.

둘째, 비상장기업으로의 적용가능성에 대해 알아보았고 이를 통해 실제 업무에서 발생하는 다양한 비상장기업의 주식가치 평가 업무에도 이를 활용할 수 있을 것으로 기대된다. 구체적으로 기업상장 업무나 기업금융 업무, 채권발행업무 등 기업 주가를 평가해야하는데 시가가 존재하지 않을 경우, 랜덤포레스트를 비롯한 머신러닝을 통해 비용과 시간 절감을 통한 실제 기업

의 업무 개선을 가능하게 할 것으로 기대된다.

한편 자연어처리를 통한 텍스트나 방송자료를 활용하지 못했다는 것은 본 연구의 한계로 남아있다. 주가 자료는 영향을 미치는 변수가 워낙 많고 태생적으로 높은 분산을 지니고 있기에 재무데이터로만 분석을 실시할 시 일정한 한계를 갖게 될 수 밖에 없다. 그러나 향후 자연어처리 기술의 적용과 함께 이를 보완하여 보다 향상된 성능으로 적정 주가를 예측하는 것이 가능해질 것으로 기대된다. 이는 향후 연구 과제로 남겨두고자 한다.

참고문헌

Kyun Sun Eo · Kun Chang Lee(2017), Predicting stock price direction by using data mining methods - Emphasis on comparing single classifiers and ensemble classifiers, 한국컴퓨터정보학회, 한국컴퓨터정보학회논문지, 22(11), 111-116

강규호(2018), "베이지안 머신 러닝을 이용한 은행권 주택담보대출 예측", 예금보험공사, 금융안정연구, 19(1), 99-129

김병수 · 오현탁(2011), "비상장기업 주식평가를 위한 결합모형의 검정", 전북대학교 산업경제연구소, Asia-Pacific Journal of Business & Commerce, 3(2), 93-112

김성진 · 안현철(2016), "기업신용등급 예측을 위한 랜덤 포레스트의 응용", 경성대학교 산업개발연구소 ,산업혁신연구, 32(1), 187-211

박근우 · 김성현 · 권혁린 · 이태규, "비상장주식 평가 실무연구", 한국공인회계사회

박종찬 · 최원석(2011), '비상장기업 주식의 가치평가 - 삼성에버랜드의 전환사채 발행사례", 한국경영학회, Korea Business Review, 15(3), 113-139

허준영 · 양진용(2015), "SVM 기반의 재무 정보를 이용한 주가 예측", 정보과학회, 컴퓨팅의 실제 논문지, 21(3), 167-172

양진용(2016), "기업 재무 정보를 활용한 머신 러닝 기반 경영 예측 시스템", 한성대학교

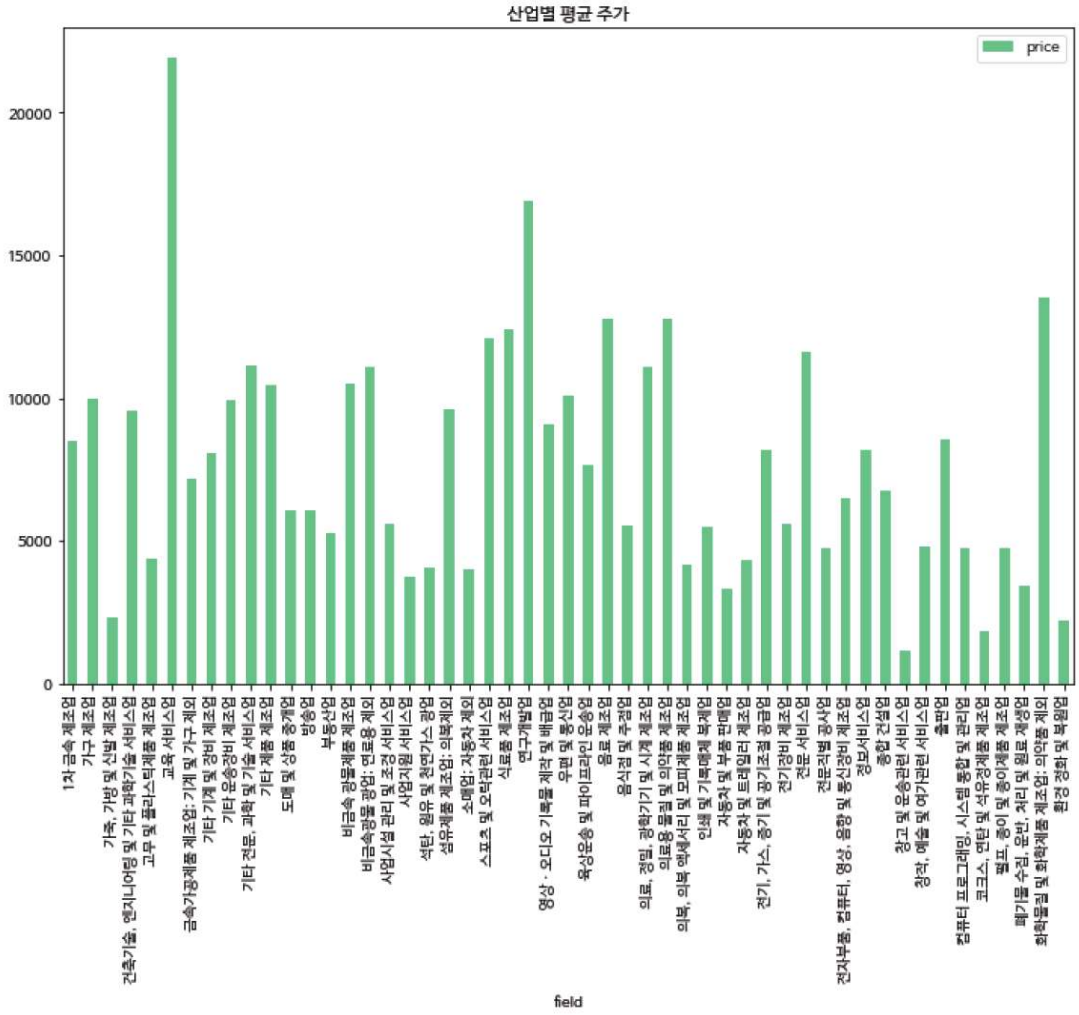
이우식(2017), “딥러닝분석과 기술적 분석 지표를 이용한 한국 코스피주가지수 방향성 예측”, 한국데이터정보과학회, 한국데이터정보과학회지, 28(2), 287-295

이지훈(2016), "딥러닝을 이용한 주가 예측 모형", 숭실대학교

<부록 1> 산업명 더미 변수 설명

전자부품, 컴퓨터, 영상, 음향 및 통신장비 제조업	3949
기타 기계 및 장비 제조업	2088
출판업	1447
의료용 물질 및 의약품 제조업	1309
도매 및 상품 중개업	1207
화학물질 및 화학제품 제조업; 의약품 제외	984
의료, 정밀, 광학기기 및 시계 제조업	838
자동차 및 트레일러 제조업	709
전기장비 제조업	665
금속가공제품 제조업; 기계 및 가구 제외	589
고무 및 플라스틱제품 제조업	521
컴퓨터 프로그래밍, 시스템 통합 및 관리업	419
1차 금속 제조업	370
영상·오디오 기록물 제작 및 배급업	288
연구개발업	268
정보서비스업	260
식료품 제조업	259
기타 운송장비 제조업	238
비금속 광물제품 제조업	210
의복, 의복 액세서리 및 모피제품 제조업	198
종합 건설업	198
펄프, 종이 및 종이제품 제조업	196
섬유제품 제조업; 의복제외	175
기타 전문, 과학 및 기술 서비스업	144
소매업; 자동차 제외	143

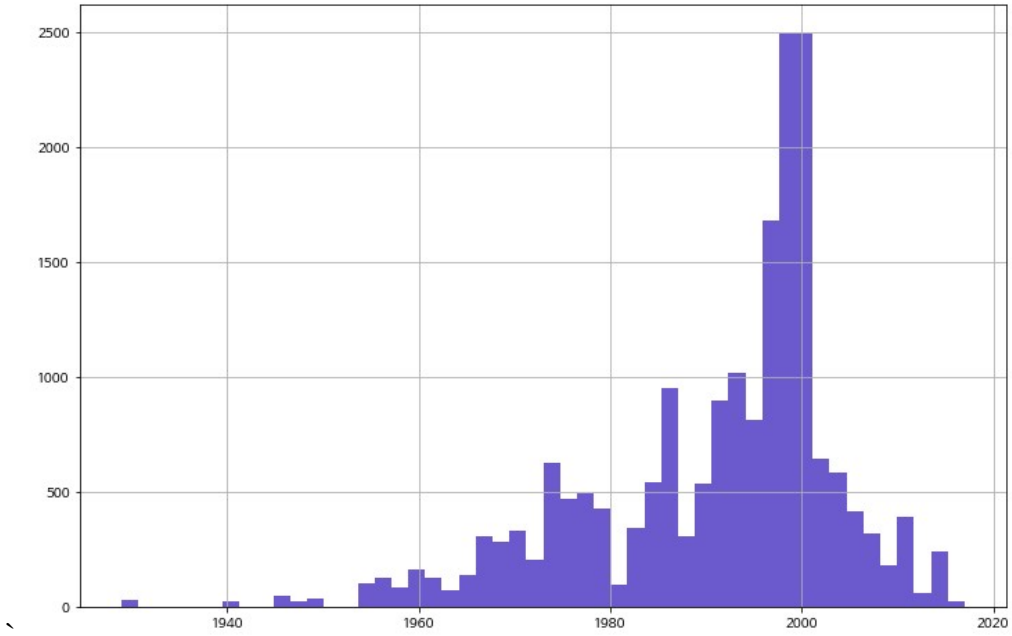
전문 서비스업	130
우편 및 통신업	127
기타 제품 제조업	103
전문직별 공사업	96
가구 제조업	96
음료 제조업	90
사업지원 서비스업	86
폐기물 수집, 운반, 처리 및 원료 재생업	86
건축기술, 엔지니어링 및 기타 과학기술 서비스업	75
인쇄 및 기록매체 복제업	71
방송업	67
부동산업	50
육상운송 및 파이프라인 운송업	46
스포츠 및 오락관련 서비스업	42
창작, 예술 및 여가관련 서비스업	34
가죽, 가방 및 신발 제조업	34
자동차 및 부품 판매업	28
창고 및 운송관련 서비스업	28
비금속광물 광업; 연료용 제외	27
석탄, 원유 및 천연가스 광업	20
환경 정화 및 복원업	18
교육 서비스업	14
코크스, 연탄 및 석유정제품 제조업	13
사업시설 관리 및 조경 서비스업	13
전기, 가스, 증기 및 공기조절 공급업	12
음식점 및 주점업	8


산업별 평균 주가
price
20000
15000
10000
5000
0
1차 금속 제조업
가구 제조업
가죽, 가방 및 신발 제조업
건축기술, 엔지니어링 및 기타 과학기술 서비스업
고무 및 플라스틱제품 제조업
교육 서비스업
금속가공제품 제조업; 기계 및 가구 제외
기타 기계 및 장비 제조업
기타 운송장비 제조업
기타 전문, 과학 및 기술 서비스업
기타 제품 제조업
도매 및 상품 중개업
발송업
부동산업
비금속 광물제품 제조업
비금속광물 광업; 연료용 제외
사업시설 관리 및 조경 서비스업
사업지원 서비스업
석탄, 원유 및 천연가스 광업
섬유제품 제조업; 의복제외
소매업; 자동차 제외
스포츠 및 오락관련 서비스업
식료품 제조업
연구개발업
영상 · 오디오 기록물 제작 및 배급업
우편 및 통신업
육상운송 및 파이프라인 운송업
음료 제조업
음식점 및 주점업
의료, 정밀, 광학기기 및 시계 제조업
의료용 물질 및 의약품 제조업
의복, 의복 액세서리 및 모피제품 제조업
인쇄 및 기록매체 복제업
자동차 및 부품 판매업
자동차 및 트레일러 제조업
전기, 가스, 증기 및 공기조절 공급업
전기장비 제조업
전문 서비스업
전문직별 공사업
전자부품, 컴퓨터, 영상, 음향 및 통신장비 제조업
정보서비스업
종합 건설업
창고 및 운송관련 서비스업
창작, 예술 및 여가관련 서비스업
출판업
컴퓨터 프로그래밍, 시스템 통합 및 관리업
코크스, 연탄 및 석유정제품 제조업
펄프, 종이 및 종이제품 제조업
폐기물 수집, 운반, 처리 및 원료 재생업
화학물질 및 화학제품 제조업; 의약품 제외
환경 정화 및 복원업
field

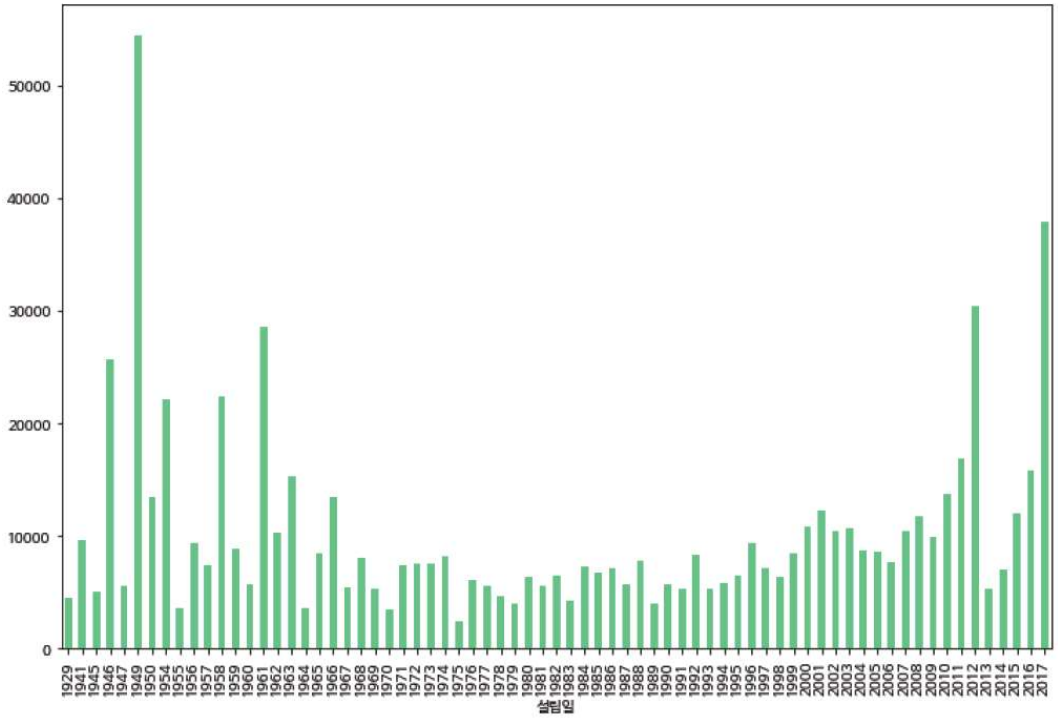
<부1-1> 산업별 평균 주가

<부록2> 기타 데이터 관련 자료

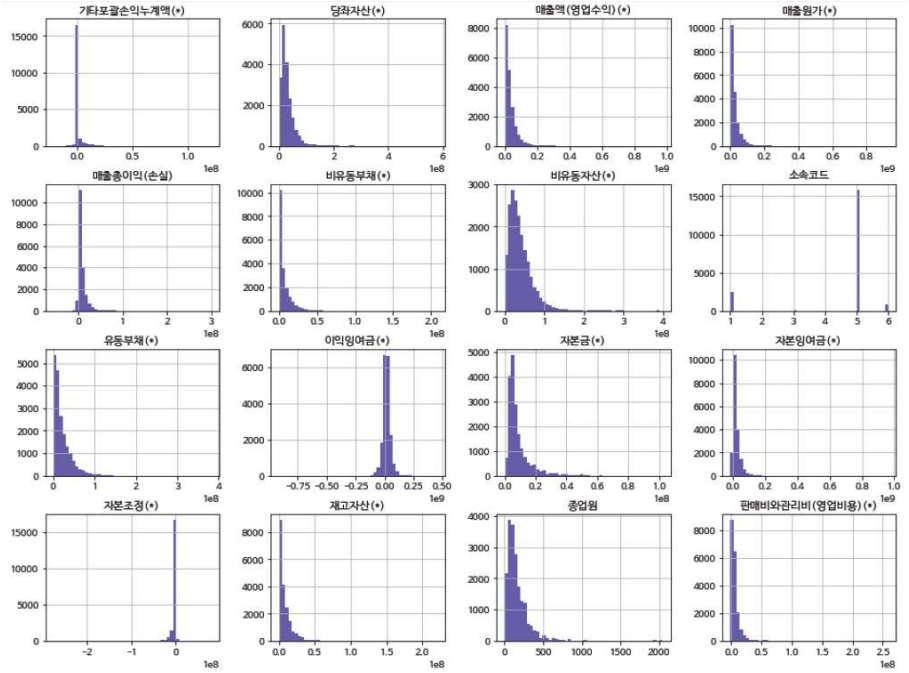
<그림 부2-1> 설립일 히스토그램



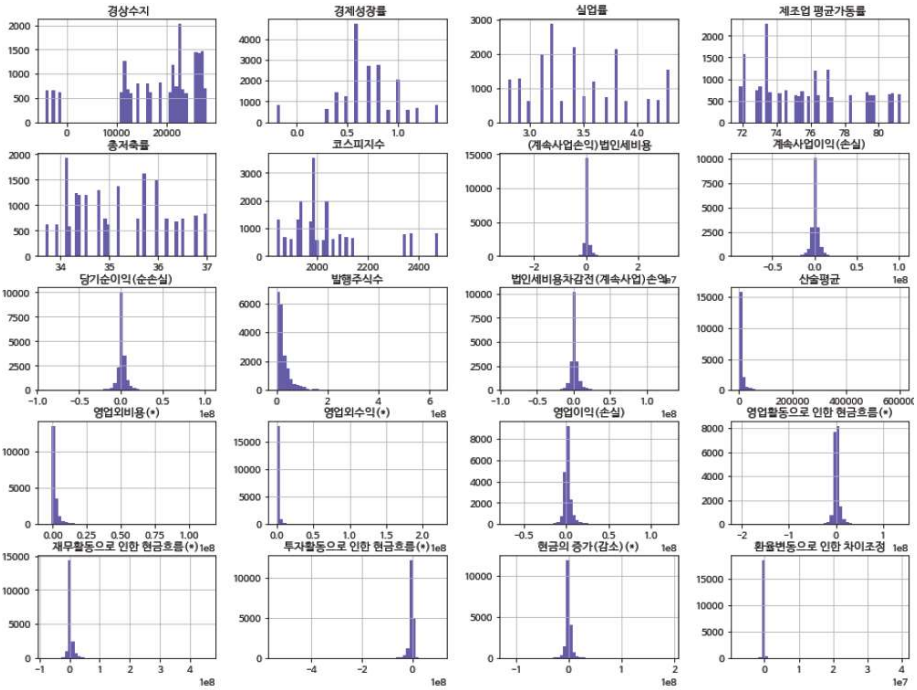
<그림 부2-2> 설립일별 평균 주가



<설명변수 히스토그램1>



<설명변수 히스토그램2>



<주요변수 기초통계량>

	종업원	당좌자산	재고자산	비유동자산	유동부채	비유동부채	자본금
count	19086	1.91E+04	1.91E+04	1.91E+04	1.91E+04	1.91E+04	1.91E+04
mean	162.263544	3.04E+07	7.87E+06	3.99E+07	2.19E+07	7.25E+06	9.19E+06
std	151.949052	2.62E+07	1.01E+07	3.28E+07	2.39E+07	1.18E+07	9.32E+06
min	3	0.00E+00	0.00E+00	0.00E+00	0.00E+00	-8.27E+03	0.00E+00
25%	70	1.43E+07	1.35E+06	1.82E+07	6.82E+06	1.10E+06	4.10E+06
50%	122	2.36E+07	4.99E+06	3.24E+07	1.43E+07	3.60E+06	6.06E+06
75%	208	3.88E+07	1.08E+07	5.27E+07	2.95E+07	9.22E+06	1.04E+07
max	2044	3.73E+08	2.20E+08	3.98E+08	3.82E+08	2.08E+08	1.02E+08
	자본조정	**기타포괄손익누계액**	**이익잉여금**	**매출액**	**매출원가**	**매출총이익**	**판관비**
count	1.91E+04	1.91E+04	1.91E+04	1.91E+04	1.91E+04	1.91E+04	1.91E+04
mean	-2.07E+06	8.67E+05	1.77E+07	3.46E+07	2.65E+07	8.03E+06	6.35E+06
std	7.07E+06	3.61E+06	3.73E+07	4.09E+07	3.48E+07	1.12E+07	8.25E+06
min	-2.76E+08	-2.11E+07	-9.11E+08	0.00E+00	0.00E+00	-5.69E+07	-1.59E+06
25%	-2.19E+06	0.00E+00	-2.77E+05	1.10E+07	7.25E+06	1.96E+06	2.05E+06
50%	-5.17E+05	0.00E+00	1.62E+07	2.30E+07	1.68E+07	4.78E+06	4.00E+06
75%	0.00E+00	3.61E+04	3.44E+07	4.45E+07	3.41E+07	1.02E+07	7.60E+06
max	8.05E+07	1.21E+08	4.51E+08	9.70E+08	9.28E+08	2.99E+08	2.59E+08

	영업이익	영업외수익	영업외비용	법인세비용차감전손익	법인세비용
count	1.91E+04	1.91E+04	1.91E+04	1.91E+04	1.91E+04
mean	1.69E+06	1.53E+06	1.91E+06	1.31E+06	3.86E+05
std	5.47E+06	3.74E+06	3.59E+06	6.98E+06	1.30E+06
min	-6.89E+07	-1.28E+05	-5.37E+06	-8.93E+07	-3.00E+07
25%	-5.26E+05	3.06E+05	3.03E+05	-8.81E+05	0.00E+00
50%	7.88E+05	7.15E+05	8.55E+05	7.42E+05	5.38E+04
75%	2.99E+06	1.62E+06	2.12E+06	3.13E+06	5.25E+05
max	1.22E+08	2.22E+08	1.14E+08	1.31E+08	3.22E+07
	계속사업이익	**당기순이익**	**영업활동현금흐름**	**투자활동현금흐름(*)**	**재무활동현금흐름**
count	1.91E+04	1.91E+04	1.91E+04	1.91E+04	1.91E+04
mean	9.20E+05	9.02E+05	1.48E+06	-3.63E+06	2.38E+06
std	6.28E+06	6.45E+06	7.08E+06	1.02E+07	1.05E+07
min	-8.93E+07	-9.01E+07	-2.06E+08	-2.68E+08	-8.06E+07
25%	-8.36E+05	-8.57E+05	-1.31E+06	-5.22E+06	-1.05E+06
50%	6.70E+05	6.66E+05	7.78E+05	-1.53E+06	0.00E+00
75%	2.70E+06	2.70E+06	3.62E+06	0.00E+00	3.34E+06
max	1.03E+08	1.03E+08	1.38E+08	1.02E+08	2.79E+08
	환율변동차이조정	**현금의 증가**	**산술평균**	**경제성장률**	**총저축률**
count	1.91E+04	1.91E+04	19086	19086	19086
mean	-6.66E+03	2.19E+05	8244.4317	0.701284	35.163382
std	4.79E+05	6.60E+06	12382.841	0.317053	0.948418
min	-7.39E+06	-1.15E+08	94	-0.2	33.7
25%	-2.67E+03	-1.67E+06	2320	0.6	34.3
50%	0.00E+00	-3.59E+04	4412	0.7	35
75%	8.10E+01	1.73E+06	8813.75	0.9	36
max	3.98E+07	1.91E+08	149179	1.4	37
	제조업 평균가동률	**실업률**	**코스피지수**	**경상수지**	**발행주식수**
count	19086	19086	19086	19086	1.91E+04
mean	75.73697	3.461998	2038.0694	18225.9855	3.02E+07
std	2.850417	0.440159	156.44897	8796.82977	4.67E+07
min	71.8	2.8	1843.47	-4253	4.08E+05
25%	73.4	3.1	1939.3	12833	1.05E+07
50%	75.4	3.4	1991.97	21472.9	1.71E+07
75%	77.3	3.8	2068.54	25944.5	3.38E+07
max	81.3	4.3	2476.37	28086.9	6.35E+08

<부록3> 머신러닝 코드

본 연구를 진행하는데 있어서 머신러닝을 활용하기 위해 파이썬을 이용하였다. 자세한 코드는 다음 깃허브에 수록하였다.

코드 수록 깃허브(Ctrl + Click) :

https://github.com/KhelKim/ForecastFairValueOfUnlistedCompany