Asia-pacific Journal of Convergent Research Interchange
Vol.9, No.7, July 31 (2023), pp.67-77
http://dx.doi.org/10.47116/apjcri.2023.07.06

A Study on the Machine Learning Model for the Financial Performance Prediction of Startups

스타트업 재무성과예측을 위한 머신러닝 모형에 관한 연구

Yong-Deok Kim[1], Hangjun Cho[2], Changhyun Lee[3], Jaemin Cho[4]

김용덕[1], 조항준[2], 이창현[3], 조재민[4]

[1] *Research Fellow, Goyang Research Institute, Korea, kyd4462@naver.com*
[2] *Deputy General Manager, DB Asset Management, Korea, chjun01@hotmail.com*
[3] *Researcher, AleconoLab Inc., Korea, dns05018@naver.com*
[4] *Associate Professor, College of Business Administration, University of Ulsan, Korea, jaemincho@ulsan.ac.kr*

Corresponding author: Jaemin Cho

Abstract: Investment decisions for startups are difficult to evaluate with the traditional valuation method of listed companies due to high risks and uncertainties caused by information asymmetry. This study analyzed samples for early startups that applied for fundraising from KOTEC (Korea technology finance corporation) and studied the most suitable model for predicting their post-investment performance using machine learning algorithms. First, Logistic regression analysis is conducted to analyze the characteristics of startups for use in machine learning algorithms, and human organizational resources and the founder's experience are found to be important variables in predicting the startup's financial performance of startups. Through this, machine learning algorithms such as SVM (Support Vector Machine), KNN (K-Nearest Neighbor) and GBM (Gradient Boosting Machine) are used to predict the financial performance of invested startups, and more than 60% of predictive power was obtained. Among the algorithms used in the analysis, GBM has a higher F1 score than other algorithms, indicating that it has better predictive power. This study has the following practical implication for angel investors or venture capital and other investment institutions. If investors have the quantitative data of startup, this study presents a suitable machine learning model for the primary screening to forecast future financial performance in the VC's investment decision-making process.

Keywords: Startups, Financial Performance, Machine Learning, Prediction Models

요약: 창업기업에 대한 투자의사결정은 높은 리스크와 정보비대칭에 의한 불확실성으로 인하여 상장기업의 전통적인 가치평가 방식으로 평가하는데 많은 어려움이 있다. 본 연구는 기술보증기금으로부터 투자 유치를 신청한 스타트업을 대상으로 표본을 정하고 분석하였으며, 머신러닝 알고리즘을 활용하여 투자이후의 성과를 예측하는데 가장 적합한 알고리즘 모형에 대하여 연구하였다. 머신러닝 알고리즘 분석을 위하여 로지스틱 회귀분석을 통하여 스타트업의 재무 성과를 예측하는데 영향을 미치는 변수들을 분석하였으며, 기업의 인적 조직 자원과 창업자의 경험 요인들을 재무성과 예측에 활용하였다. 이를 통해, SVM(Support Vector

Received: March 11, 2023; 1st Review Result: April 26, 2023; 2nd Review Result: May 24, 2023
Accepted: June 30, 2023

Machine), KNN(K-Nearest Neighbor), GBM(Gradient Boosting Machine) 등의 머신러닝 알고리즘을 사용하였을 때 60% 수준 이상의 예측력을 얻을 수 있었으며, 분석에 사용된 알고리즘 중에서는 GBM이 다른 알고리즘보다 F1 score가 더 높아 보다 우수한 예측력을 갖고 있는 것으로 나타났다. 본 연구는 엔젤투자자 또는 벤처캐피털 및 기타 투자기관 등이 스타트업의 초기 정량적 데이터를 가지고 있을 경우, 투자 의사결정 과정에서 향후 재무 성과를 예측하기 위한 1차 스크리닝에 적합한 머신러닝 모형으로 활용할 수 있다는 점에서 실무적 의의를 가지고 있다.

핵심어: 스타트업, 투자성과, 머신러닝, 예측모형

1. 서론

혁신적인 기업으로 성장할만한 스타트업을 초기에 선별하는 일은 투자 유관기관뿐만 아니라 정부, 업계 관계자, 구직자와 같은 다양한 이해 관계자들도 관심을 갖는 분야이다. 특히, 창업기업에 대한 초기투자의사결정은 높은 리스크와 정보비대칭에 의한 불확실성으로 기존의 상장기업과 비슷한 방식으로 평가하고 투자결정을 하는데 많은 어려움이 있다. 그동안 스타트업 성과와 관련된 선행연구 중에서 주요 투자 주체인 엔젤투자자와 벤처캐피탈들과 관련된 연구를 분석해보면, 이들이 투자결정과정에서 가장 중점적으로 분석하는 주요 공통 요인은 상품 및 서비스, 시장, 투자 수익률, 창업가의 역량 크게 4가지로 요약된다[1]. 그 중에서도 스타트업은 대부분의 초기 자원이 창업자의 역량에 크게 영향을 받기 때문에 창업가의 사회적, 기술적, 기업가적 특성에 의해 기업의 성과가 결정되는 경우가 많다[2][3].

특히, 벤처캐피탈은 엔젤투자자와 액셀러레이터보다 시리즈 A 단계 이상의 성장한 기업에 투자하는 경우가 대부분이다. 따라서, 대표자의 학력, 경력 등의 기술 및 경영 전문성, 기업가적 특성 등을 함께 고려하여 투자의사결정을 하는데, 이때는 기업실사와 투자심의위원회 등과 같은 검증 과정을 통하여 수익률, 지분현황 등에 대한 검토를 충분히 거친다[4][5]. 즉, 초기의 높은 불확실성을 감내하고 기대수익률을 바라는 스타트업 투자의 특성으로 인하여 우수한 성과가 예측되는 기업에 대한 실사과정으로 많은 시간과 자원이 투입되고 있다. 이에 본 연구는 실제 투자 유치에 성공한 초기 스타트업을 대상으로 창업자의 기업가적 특성을 정형화하여 재무성과를 분석하였으며, 머신러닝 알고리즘을 활용하여 투자이후의 재무 성과를 예측하는데 가장 적합한 머신러닝 모형에 대하여 연구하였다.

본 연구의 목차는 다음과 같다. 서론에 이어 2장에서는 머신러닝의 정의와 관련 선행연구를 살펴보았으며, 제 3장에서는 연구모형을 설명한다. 제 4장에서는 실증분석 결과를 제시하며, 제 5장에서 결론을 기술한다.

2. 이론적 배경

2.1 머신러닝(machine learning)의 정의

머신러닝(machine learning)은 인공지능의 한 분야로서 주어진 데이터 분석을 위해 컴퓨터가 스스로 학습할 수 있도록 알고리즘을 개발하여, 데이터가 가진 패턴을

찾아내는 기법을 의미하며, 예측 정확도가 높을 뿐만 아니라, 예측 시간을 단축하고 자동화된 대응을 수행하므로 여러 분야에서 활용되고 있다[6].

머신러닝(Machine Learning)에 대한 정의는 크게 두 가지로 볼 수 있는데, Samuel[7]의 '컴퓨터가 명시적인 프로그래밍 없이 학습할 수 있도록 하는 연구 분야'라는 정의와, Mitchell[8]의 '만약 임의의 작업(T)과 관련된 경험(E)을 통해 성능 측정 방법(P)으로 성능을 측정했을 때 성능이 향상된다면 이러한 컴퓨터 프로그램은 학습을 한다고 말할 수 있다'라는 정의이다. 최근의 머신러닝은 대부분 후자의 정의를 따른다. 머신러닝의 학습 방법은 학습 데이터(training data)에 레이블(label)이 있는 경우와 있지 않은 경우로 나누어 학습방법을 구분하는데, 크게 지도 학습(Supervised Learning), 비지도 학습(Unsupervised Learning) 그리고 강화 학습(Reinforcement Learning)이 있다.

지도학습은 입력값과 그에 따른 결과값이 주어진 상태에서 학습하는 방식을 의미하고, 분류(Classification)나 회귀분석이 지도학습에 속한다. 비지도학습은 결과값이 없는 상태에서 입력값만으로 학습시키는 방식으로, 군집화(Clustering)나 차원축소가 이에 해당한다. 비지도 학습 머신러닝은 데이터 구조와 무엇을 나타내는지를 미리 알지 못한 상태로 데이터를 표현하는데, 지도학습 머신러닝처럼 데이터와 분류 결과 사이의 매핑함수를 만드는 목적은 동일하다. 정확한 정답을 알 때까지 반복적으로 추가 데이터를 기반으로 매핑 함수를 정제하고 재조정한다. 비지도 학습 머신러닝의 예로는 기댓값-최대화(EM, Expectation-Maximization) 알고리즘과 K-평균 클러스터링(K-means Clustering) 알고리즘이 있다. 비지도 학습 머신러닝은 데이터를 랜덤으로 추출하여 처리하여 구분하고, 본 알고리즘에 의해 입력 이미지가 새롭게 입력되면 기존의 학습된 범주의 판별자로부터 받은 결과 값을 가지고 자율적으로 판별한다[9]. 강화학습은 현재 상태에서 행동에 따른 보상이 주어질 때 이를 최대화하도록 학습하는 방식을 의미한다. 강화학습은 지도학습에 포함시키기도 하고, 머신러닝의 또 다른 종류로 구별짓기도 하는데, 알고리즘이 수행한 결과에 따라 수행 방식을 진화시켜 나가는 알고리즘을 의미한다[10].

2.2 머신러닝(machine learning)을 이용한 예측(Prediction)연구

머신러닝의 개념을 Samuel[7]이 정립한 이후, 현재 여러 산업 및 학문 분야에서 다양하게 활용되고 있다. 특히, 과거에 비해 끊임없이 생성되는 대용량의 정형/비정형 데이터 수집으로 인한 빅 데이터로 인하여 예측이 중요한 경영 및 경제 분야에서의 머신러닝 활용 수준은 점점 높아지고 있다[11]. 특히, 머신러닝을 활용할 경우 다양한 유형의 데이터 분석이 가능하고, 더 많은 양의 데이터를 한 번에 분석할 수 있을 뿐만 아니라, 모형의 유연성 확보로 더욱 정교한 특징을 추출이 가능하며, 예측력이 높다는 강점이 있다. 국내 연구를 살펴보면, 라윤선, 최홍식, 김선웅[12]은 SVM을 이용해 VKOSPI 변화를 예측하여 옵션 매매에 적용하였고, 김명현, 이세호, 신동훈[13]은 K-Nearest Neighbors(KNN) 알고리즘을 통해 Kospi200 지수 선물 가격을 예측하였다. 또한, 박석진, 정재식[14]은 머신러닝 모형을 이용하여 KOSPI200 선물가격의 상승과 하락 방향을 예측하였으며, 윤종문[15]은 딥러닝 알고리즘을 이용하여 신용카드 부도 예측 모형을 추정하였다. 강규호[16]는 베이지안 머신러닝 알고리즘을 이용하여 주택담보대출 규모를 예측하는 것에 더해 향후 증가율까지 전망하였다. 또한, 황윤태[17]는 아파트 실거래가 데이터를 이용하여 다양한 머신러닝 알고리즘으로 예측모형을 설계하여

비교해보기도 하였다.

여기서 주로 쓰이는 머신러닝 방법론은 SVM(Support Vector Machine), 토픽모델(topic model), 딥러닝(deep learning), 트리 앙상블(tree ensembles), 인과 포레스트(causal forest), 뉴럴 네트워크 임베딩(neural network embedding), 능동적 학습(active learning), 강화학습(reinforcement learning) 등이 있다. 머신러닝을 활용한 연구는 모델의 활용 목적에 따라 예측(prediction), 특징 추출(feature extraction), 기술적 해석(descriptive interpretation), 인과적 해석(causal interpretation), 처방 분석(prescription analysis), 최적화·추정(optimization·estimation)이 있다. 주로 특징 추출, 기술적 해석, 예측을 하는데 많이 쓰이는데 반해, 인과적 해석이나 처방분석은 부족한 상황이다. 이렇듯 경영학의 다양한 연구분야에서 머신러닝 기법을 활용한 연구가 활발하게 진행되고 있으며, 향후 활용 분야는 더욱 확대될 것으로 생각된다.

3. 연구 설계

3.1 연구 모형

본 연구에서 사용한 데이터에는 스타트업의 성과라는 결과값이 주어져 있기 때문에 지도학습에 해당하는 머신러닝 알고리즘들을 사용하였다. 지도학습에 속하는 머신러닝 알고리즘으로는 대표적으로 회귀분석, SVM, KNN, 의사결정나무, 랜덤포레스트, GBM 등이 있다. 회귀분석의 경우 일반적인 OLS뿐만 아니라 범주형 종속변수를 위한 로지스틱 회귀분석과 함께 다항회귀분석에서의 예측력과 설명력을 높이기 위해 정규화를 도입한 Lasso 회귀분석과 Ridge 회귀분석 등도 머신러닝 알고리즘 범주에 포함된다. Cortes and Vapnik[18]의 SVM은 주어진 데이터들을 고차원의 공간에 사상시킨 후 이 데이터들을 가장 잘 분류하는 경계를 찾아내는 방식이다. Fix and Hodges[19]에 의해 처음 소개된 KNN은 임의의 데이터를 분류할 때 가장 근접한 K개 이웃 데이터를 기반으로 분류하는 방식이다. 의사결정나무는 입력 데이터와 결과값과의 관계를 결정나무의 형태로 표현하는 방식이며, 랜덤포레스트는 의사결정나무를 여러 개 만들어 학습시키고 각 결정나무에서 나온 결과를 종합하여 일반화시키는 방식이다. GBM은 Friedman[20]로부터 시작된 것으로, 단일 의사결정나무로부터 출발하여 예측 오차를 줄여나가는 방향으로 학습해나가는 방식이다. 본 연구에서는 투자 이후 성과는 재무적인 관점에서 ROA, ROE를 기준으로 로지스틱 회귀분석, SVM, KNN, 랜덤포레스트, GBM을 사용하였다.

3.2 분석 자료

본 연구는 2013년부터 2017년까지 기술보증기금의 투자를 받은 기업들의 데이터를 기반으로 한다. 중소기업창업지원법 상의 창업기업에 대한 정의에 따라 창업년도로부터 7년이 경과하지 않은 기업을 대상으로 하였으며, 기술보증기금의 투자 직전과 투자 이후의 재무제표 데이터가 존재하여 정량적 비교가 가능한 기업들의 데이터를 사용하였다. 이에 따라 분석에 선택된 기업의 수는 112개이다. 분석에 사용한 기업 특성과 관련된 데이터는 대표자연령, 대표자학력, 대표자동업종사년, 경영진수, 지분보유경영진수, 주요기술인력수, 상시종업원수, 기술개발전담조직, 특허건수, 특허출원중건수, 실용신안건수, 기술평가등급으로 총 12개이고, 대표자학력, 기술개발전담조직, 기술평가등급과 같이 범주화는 되어 있으나 숫자로 표현되지 않는

데이터들은 5점 척도(최저 1점 ~ 최고 5점)로 변환하여 사용하였다. 여기에서 대표자학력의 경우에는 대학원박사=5, 대학원석사=4, 대학교=3, 전문대=2, 고등학교이하=1으로 구분하였으며, 기술개발전담조직의 경우에는 기업부설연구소=5, 연구개발전담부서=4, 기술개발연구실=3, 기술개발인력만 확보=2, 해당없음=1이며, 기술평가등급의 경우 (TI2, AAA, AA)=5, (BBB, TI3)=4, (BB, TI4)=3, (B, TI5)=2, D=1로 변환하였다. 스타트업의 성과와 관련된 지표는 당기순이익을 총자산으로 나눈 총자산이익률(ROA), 당기순이익을 자기자본으로 나눈 자기자본이익률(ROE)을 사용하였고, 기술보증기금의 투자 전후 스타트업 성과 지표의 증가여부(증가했을 때 1, 그렇지 않을 때 0으로 부여)를 머신러닝 알고리즘에서의 학습대상 변수로 활용하였다. 즉, ROA 증가의 경우 투자 받은 시점을 T라고 할 때 ROA(T) - ROA(T-1)로 정의하였고, ROE 증가도 유사하게 ROE(T) - ROE(T-1)로 정의하였다. 분석 데이터의 기술 통계량은 아래 [표 1]과 같다.

[표 1] 분석 데이터 기술 통계량

[Table 1] Descriptive Statistics

변수명	평균	표준편차	최소값	중간값	최대값	관측수
ROA 증가*	0.12	2.34	-1.53	-0.05	24.39	112
ROE 증가*	-1.03	5.84	-29.11	-0.27	20.66	112
대표자연령	49.98	7.37	33	49	72	112
대표자학력*	3.34	1.11	1	3	5	112
대표자동업종사년	15.30	8.14	0	15	38	112
경영진수	1.83	1.08	0	2	5	112
지분보유경영진수	0.76	0.86	0	1	3	112
주요기술인력수	7.21	4.02	2	6	26	112
상시종업원수	28.95	29.73	0	17	132	112
기술개발전담조직*	4.71	0.69	2	5	5	112
특허건수	2.77	6.97	0	1	65	112
특허출원중건수	1.82	4.22	0	0	38	112
실용신안건수	0.05	0.30	0	0	2	112
기술평가등급*	4.26	0.74	2	4	5	112

4. 분석 결과

4.1 로지스틱 회귀분석

먼저, 스타트업의 재무적 성과를 예측하기 위해 ROA 및 ROE 증가여부를 종속변수로, 스타트업의 기업가적 특성을 독립변수로 하여 로지스틱분석을 하였으며, 결과는 [표 2]와 같다.

ROA를 스타트업 성과 지표로 간주하여 분석한 결과, 투자를 받은 후 우수한 성과를 내는 스타트업은 대표자연령이 상대적으로 낮고 경영진수가 많은 경향을 보이는데,

대표자연령의 경우 5% 수준에서 유의하고 경영진수의 경우 10% 수준에서 유의한 것으로 나타났다. 이는 ROE를 스타트업 성과 지표로 간주하여 분석한 결과와 동일하게 나타났다. 이 결과는 기존의 창업기업 및 중소기업의 성과에 있어서 대표자학력, 대표자동업종사년 등 경영진의 역량이 중요한 것으로 나타난 기존의 연구[21][22]와는 상반된다. ROA와 ROE를 성과 지표로 간주하여 분석한 결과의 R^2를 보면 각각 8.86%, 10.02% 수준으로 낮게 나타났다. 본 연구에서는 로지스틱 모형을 이용하여 스타트업 재무성과에 영향을 미치는 변수들을 먼저 분석한 후, 머신러닝 알고리즘을 이용한 분석을 시도한다.

[표 2] 로지스틱 회귀분석 결과

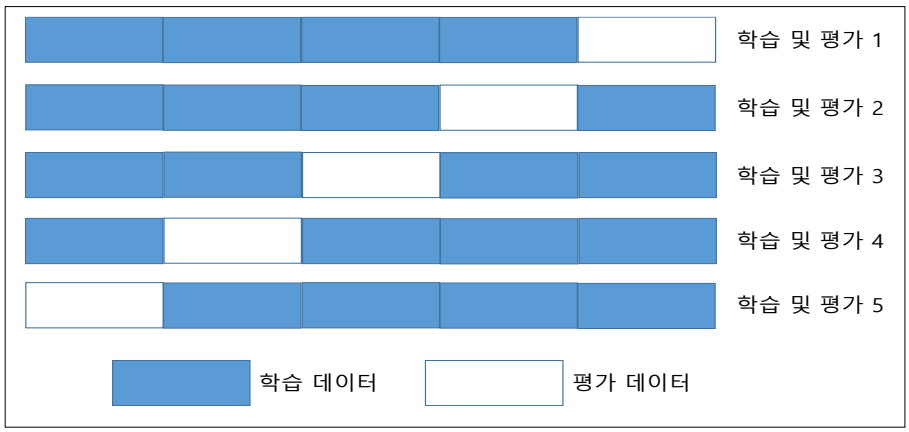
[Table 2] Logistic Regression Anlaysis

ROA 증가여부			ROE 증가여부		
변수	계수	z값	변수	계수	z값
상수	2.578	1.059	상수	-1.329	-0.531
대표자연령	-0.086**	-2.146	대표자연령	-0.076**	-2.020
대표자학력	0.003	0.012	대표자학력	0.073	0.303
대표자동업종사년	0.013	0.363	대표자동업종사년	-0.017	-0.512
경영진수	0.410*	1.652	경영진수	0.405*	1.657
지분보유경영진수	-0.285	-0.958	지분보유경영진수	-0.426	-1.451
주요기술인력수	-0.113	-1.318	주요기술인력수	-0.090	-1.164
상시종업원수	0.003	0.260	상시종업원수	0.005	0.484
기술개발전담조직	-0.015	-0.045	기술개발전담조직	0.640	1.569
특허건수	-0.054	-0.776	특허건수	-0.110	-1.403
특허출원중건수	0.029	0.540	특허출원중건수	0.039	0.724
실용신안건수	-0.292	-0.330	실용신안건수	0.091	0.129
기술평가등급	0.221	0.621	기술평가등급	0.386	1.099
R^2	8.86%		R^2	10.02%	

*, **은 각각 10%, 5% 수준에서 유의함을 의미한다.

4.2 머신러닝 알고리즘을 활용한 분석

스타트업으로의 투자 이후 우수한 성과를 예측하기 위한 모형을 분석하기 위해 머신러닝 알고리즘(SVM, KNN, 랜덤포레스트)을 사용하였다. 분석의 강건성을 높이기 위해 교차검증 방식으로 진행하였다. 교차검증은 아래 [그림 1]과 같이 전체 데이터를 k개로 나누어 학습과 정확도 평가를 k번 반복하는 방식으로, 정확도 평가에 사용되는 데이터가 중복되지 않도록 바꾸면서 평가를 진행하고 각 평가에서 얻은 정확도의 평균값을 해당 머신러닝 알고리즘의 평균적인 정확도로 간주한다. 본 연구에서는 그림과 같이 k=5로 설정하여 교차검증을 실시하였다.



[그림 1] 교차검증 예시

[Fig. 1] Example of Cross-validation

아래 [표 3]은 ROA 증가여부에 대한 머신러닝 알고리즘별 예측 결과이다. 전반적으로 60%를 상회하는 정확도를 보였고, SVM을 활용하여 스타트업의 성과와 기업 특성 사이의 관계를 예측했을 때 평균적으로 68.74%의 정확도를 보였으며, KNN의 경우 70.55%, 랜덤포레스트의 경우 66.05%, GBM의 경우 59.72%의 정확도를 보였다. 다만, 머신러닝 알고리즘 전반적으로 ROA가 증가하는 것으로 예측한 경우보다 감소하는 것으로 예측한 경우가 더 많았다. 이는 실제로 약 70% 정도의 스타트업 성과가 투자 이후 부진해졌기 때문에 머신러닝 알고리즘을 통해 학습하고 예측했을 때 과적합되는 경향이 있는 것으로 보인다.

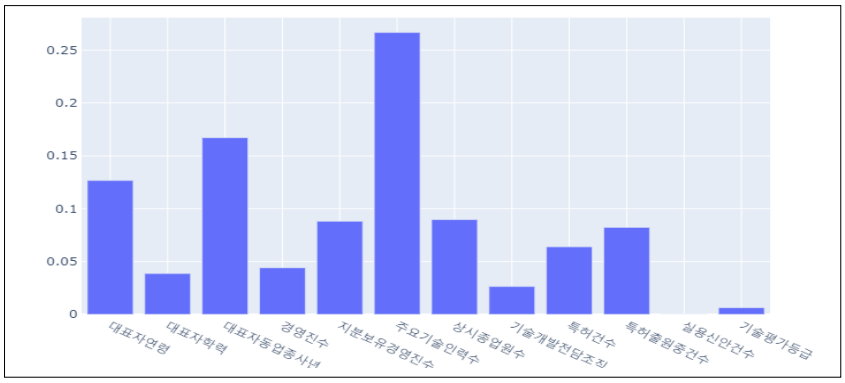
[표 3]에서의 F1 score는 정밀도(실제로 증가한 데이터를 머신러닝 알고리즘이 증가로 예측한 비율)와 재현율(머신러닝 알고리즘이 증가로 예측한 데이터 중에서 실제로 증가한 데이터의 비율)의 조화평균 값으로 계산되는데, 이 값이 클수록 머신러닝 알고리즘을 통해 예측한 결과가 증가 혹은 감소로 치우치지 않고 적절하게 예측함을 의미하고, 정확도와 함께 머신러닝 알고리즘의 예측 결과를 평가하는 지표로 사용된다. ROA 증가여부를 예측한 4개의 머신러닝 알고리즘 중에서는 GBM의 F1 score가 21.05%로 가장 높은 값을 나타냈다.

[표 3] ROA 증가에 대한 머신러닝 알고리즘 예측 결과

[Table 3] Machine Learning Algorithm Prediction Results for ROA Increase

SVM			KNN			랜덤포레스트			GBM		
예측 / 실제	감소	증가	예측 / 실제	감소	증가	예측 / 실제	감소	증가	예측 / 실제	감소	증가
감소	73	5	감소	66	7	감소	72	6	감소	61	17
증가	30	4	증가	36	3	증가	32	2	증가	28	6
평균 정확도	68.74%		평균 정확도	61.70%		평균 정확도	66.05%		평균 정확도	59.72%	
F1 score	18.60%		F1 score	12.24%		F1 score	9.52%		F1 score	21.05%	

[그림 2]는 GBM 알고리즘을 사용하여 스타트업의 ROA 성과를 예측할 때 각 스타트업 특성들의 중요도를 나타낸 것이다. 가장 중요도가 높은 스타트업의 특성으로는 주요기술인력수, 대표자동업종사년, 대표자연령 순으로 나타났다. [그림 2]를 통해 ROA 관점에서 투자를 받은 스타트업의 향후 성과를 예측하기 위해서는 스타트업의 인적 자원(주요기술인력수)과 대표자의 경험적 능력(대표자동업종사년, 대표자연령)을 주요 요인으로 고려해야 함을 알 수 있다. 한편, 로지스틱 분석 결과에서 성과를 예측하는데 대표자연령과 경영진의 수가 통계적으로 유의성이 있는 것으로 나타났는데, 이는 GBM 알고리즘 결과와도 일부 유사성이 가진다고 볼 수 있다.



[그림 2] ROA 예측시 스타트업 특성의 중요도(GBM)

[Fig. 2] Importance of Startup Characteristics in ROA Prediction (GBM)

아래 [표 4]는 ROE 증가여부에 대한 머신러닝 알고리즘별 예측 결과이다. 머신러닝 알고리즘 전반적으로 60% 수준의 정확도를 보였고, 모형별로 SVM은 66.13%, KNN은 67.83%, 랜덤포레스트는 57.15%, GBM은 57.94%의 평균 정확도를 기록하였다. [표 3]의 예측 결과와 유사하게 ROE가 증가하는 것으로 예측하는 경우보다 감소하는 것으로 예측하는 경우가 더 많았는데, ROE의 경우에도 스타트업이 투자를 받은 후에 감소하는 경우가 약 65%로 더 많은 비중을 차지하기 때문에 머신러닝 알고리즘이 과적합되는 경향이 있는 것으로 보인다.

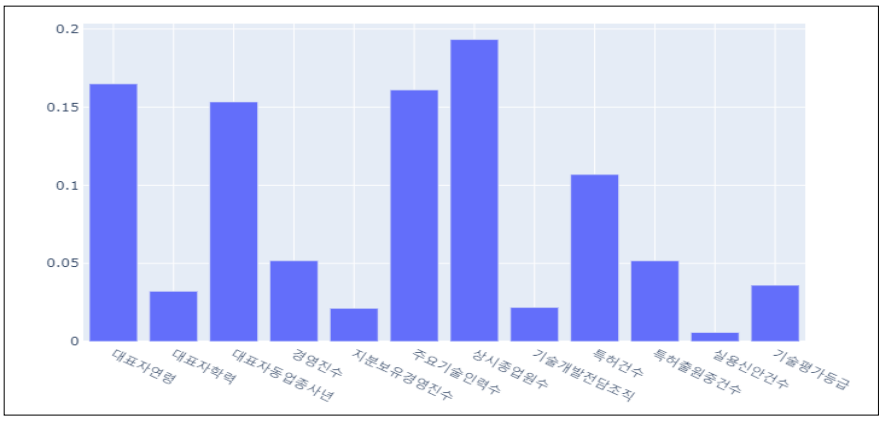
[표 4] ROE 증가에 대한 머신러닝 알고리즘 예측 결과

[Table 4] Machine Learning Algorithm Prediction Results for ROE Increase

SVM			KNN			랜덤포레스트			GBM		
예측 / 실제	감소	증가	예측 / 실제	감소	증가	예측 / 실제	감소	증가	예측 / 실제	감소	증가
감소	72	1	감소	73	0	감소	61	12	감소	51	22
증가	37	2	증가	36	3	증가	36	3	증가	25	14
평균 정확도	66.13%		평균 정확도	67.83%		평균 정확도	57.15%		평균 정확도	57.94%	
F1 score	9.52%		F1 score	14.29%		F1 score	11.11%		F1 score	37.33%	

[표 4]에서의 F1 score는 SVM이 9.52%, KNN이 14.29%, 랜덤포레스트가 11.11%, GBM이 37.33%를 기록했고, [표 3]와 마찬가지로 GBM이 가장 높은 F1 score 값을 나타냈다. 이는 GBM이 다른 알고리즘들에 비해 한 쪽으로 쏠리지 않는 예측 결과를 낼 수 있음을 보여준다.

[그림 3]은 GBM 알고리즘을 사용하여 스타트업의 ROE 성과를 예측할 때 각 스타트업 특성들의 중요도를 나타낸 것이다. 가장 중요도가 높은 스타트업의 특성으로는 상시종업원수, 대표자연령, 주요기술인력수, 대표자동업종사년, 특허건수 순으로 나타났다. 앞서 로지스틱 회귀분석에서 대표자연령만이 스타트업의 성과를 예측하는데 통계적으로 유의성이 있는 것으로 나타난 결과와 일부 유사성이 있는 것으로 보인다. [그림 3]을 통해서 ROE 관점에서 투자를 받은 스타트업의 향후 성과를 예측하기 위해서는 스타트업의 인적 자원(상시종업원수, 주요기술인력수)과 더불어 대표자의 경험적 능력(대표자연령, 대표자동업종사년), 기술적 능력(특허건수)을 주요 요인으로 고려해야 함을 알 수 있다.



[그림 3] ROE 예측시 스타트업 특성의 중요도(GBM)

[Fig. 3] Importance of Startup Characteristics in ROE Prediction (GBM)

5. 결론

본 연구는 로지스틱 회귀분석, SVM, KNN, 랜덤포레스트, GBM 등의 머신러닝 알고리즘을 활용하여 창업일로부터 7년이 경과하지 않은 창업기업의 투자 이후 성과를 예측하였다. 투자 이후 성과는 재무적인 관점에서 ROA, ROE를 기준으로 분석하였다. 연구로부터 도출된 결과는 다음과 같다. 먼저, 머신러닝 알고리즘에 사용된 스타트업의 특성들 중에서 스타트업의 성과를 예측하는데 중요한 역할을 하는 변수로는 스타트업의 인적 자원과 대표자의 경험적 능력인 것으로 나타났다. 세부적으로 보면, ROA 관점에서 투자를 받은 스타트업의 향후 성과를 예측하기 위해서는 스타트업의 인적 자원(주요기술인력수)과 대표자의 경험적 능력(대표자동업종사년, 대표자연령)을 주요 요인으로 고려해야 함을 알 수 있다. ROE 관점에서 투자를 받은 스타트업의 향후 성과를 예측하기 위해서는 스타트업의 인적 자원(상시종업원수, 주요기술인력수)과 더불어 대표자의 경험적 능력(대표자연령, 대표자동업종사년), 기술적 능력(특허건수)을 주요 요인으로 고려해야 함을 알 수 있다. 일반적으로 투자를 받기 위해 중요한

요인으로 고려되는 대표자의 학력은 투자 이후의 우수한 성과에는 중요한 영향을 미치지 않는 것으로 보인다.

투자를 받은 스타트업의 우수한 성과를 예측하기 위해서 머신러닝 알고리즘을 사용하였을 때 60% 수준 이상의 예측력을 얻을 수 있었다. 분석에 사용된 알고리즘 중에서는 GBM이 다른 알고리즘보다 F1 score가 더 높아 보다 우수한 예측력을 갖고 있는 것으로 나타났다. 실제로 투자를 받은 스타트업의 재무적 성과가 개선된 기업들보다 그렇지 않은 기업들이 더 많았기 때문에 GBM을 제외한 다른 알고리즘들은 예측 결과가 한 쪽으로 쏠리는 모습을 보였다. 즉, 투자를 받은 스타트업의 우수한 성과를 예측하는데 상대적으로 우수한 머신러닝 알고리즘은 GBM이라고 볼 수 있다.

향후에는 투자를 받은 스타트업 데이터를 더 많이 확보하고, 더 많은 변수를 추가하여 분석하면 더욱 강건한 우수 성과 스타트업 예측 모델을 구축할 수 있을 것이다. 또한, 본 연구에서 사용되지 않은 다른 알고리즘 모형을 활용하여 예측 모델을 만들고 본 연구에서의 결과와 비교하면 더욱 정확도가 높은 우수 성과 스타트업 예측 모델을 만들 수 있을 것으로 기대된다.

References

[1] A. L. Maxwell, S. A. Jeffrey, M. Lévesque, Business angel early stage decision making, Journal of Business Venturing, (2011), Vol.26, No.2, pp.212-225.
DOI: https://doi.org/10.1016/j.jbusvent.2009.09.002

[2] N. Bosma, M. Van Praag, R. Thurik, G. De Wit, The value of human and social capital investments for the business performance of startups, Small Business Economics, (2004), Vol.23, No.3, pp.227-236.
DOI: http://dx.doi.org/10.47116/apjcri.2022.08.0X

[3] Man-Goo Kang, Jae-Hyoung Kim, Min-Chan Pyo, The Effect of Entrepreneurship on Entrepreneurial Intention to Start a Start-Up : Focused on Mediating Effects of Global Mindset, Korean Jouranl of Business Administration, (2018), Vol.31, No.5, pp.1033-1054.
DOI: http://doi.org/10.18032/kaaba.2018.31.5.1033

[4] Bahn Sung-Sik, Song Kyoung-Mo, Evaluation Model and Factors for Investment Decision Making of Korean Venture Capitalists, Korean Jouranl of Business Administration, (2004), Vol.42, pp.267-291.

[5] K. Koo, C. Lee, W. Yoo, Academic research: Deriving investment decision models for Korean venture capital and analyzing their validity, Journal of Small Business and Entrepreneurship (JSBI), (2008), Vol.11, No.4, pp.1-20.

[6] Younghye Bae, Jongsung Kim, Wonjoon Wang, Younghoon Yoo, Jaewon Jung, Hung Soo Kim, Monthly Inflow Forecasting of Soyang River Dam Using VARMA and Machine Learning Models, Journal of Climate Research, (2019), Vol.14, No.3, pp.183-198.
DOI: https://doi.org/10.14383/cri.2019.14.3.183

[7] A. L. Samuel, Some Studies in Machine Learning Using the Game of Checkers, IBM Journal of Research and Development, (1959), Vol.3, pp.210-229.

[8] T. Mitchell, Machine Learning, McGraw Hill, (1997)

[9] Sang-Moo Park, Seong-Jin Kim, Dong-Hyung Lee, Soo-Dong Lee, Cheol-Young Ock, A Study on Unsupervised Learning Method of RAM-based Neural Net, Journal of the Korea Society of Computer and Information, (2011), Vol.16, No.1, pp.31-38.
UCI: G704-001619.2011.16.1.018

[10] M. Oh, H. Choi, S. Kim, J. Jang, H. Chang, M. Chun, Big data analysis and predictive model research based on machine learning for social security system in Korea, Korea Institute for Health and Social Affairs, (2017)

[11] J. Kleinberg, J. Ludwig, S. Mullainathan, Z. Obermeyer, Prediction policy problems. American Economic Review, (2015), Vol.105, No.5, pp.491-95.
DOI: https://doi.org/10.1257/aer.p20151023

[12] Y. S. Ra, H. S. Choi, S. W. Kim, VKOSPI Forecasting and Option Trading Application Using SVM, Journal of Intelligence and Information Systems, (2016), Vol.22, No.4, pp.177-192.
DOI: https://doi.org/10.13088/jiis.2016.22.4.177

[13] Myeong-Hyeon Kim, Seho Lee, Dong-hoon Shin, Predictability Test of K-Nearest Neighbors (K-NN) Algorithm : Application to the KOSPI 200 Futures, Korean Journal of Business Administration, (2015), Vol.28, No.10, pp. 2613-2633.
DOI: https://doi.org/10.18032/kaaba.2015.28.10.2613

[14] S. Park, J. Jung, Forecasting ability of machine learning algorithms using high-frequency data: KOSPI200 futures, Journal of Money and Finance, (2019), Vol.33, No.4, pp.31-60.
DOI: https://dx.doi.org/10.21023/JMF.33.4.2

[15] J. Yoon, Effectiveness Analysis of Credit Card Default Risk with Deep Learning Neural Network, Journal of Money and Finance, (2019), Vol.33, No.1, pp.151-183.
DOI: https://dx.doi.org/10.21023/JMF.33.1.5

[16] Kang Kyu Ho, Mortgage Loan Prediction: Bayesian Machine Learning Approach, Financial Stability Studies (FSS), (2018), Vol.19, No.1, pp.99-129.
DOI: https://dx.doi.org/10.26588/kdic.2018.19.1.004

[17] Yoon Tae Hwang, A Study on the Estimation of Apartment Price Index: Focused on the Machine Learning Algorithm, Journal of Money and Finance, (2019), Vol.33, No.3, pp.51-83.
DOI: https://dx.doi.org/10.21023/JMF.33.3.3

[18] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning, (1995), Vol.20, No.3, pp.273-297.

[19] E. Fix, J. L. Hodges, Discriminatory analysis, nonparametric discrimination, (1951)

[20] J. H. Friedman, Greedy Function Approximation: A Gradient Boosting Machine, IMS 1999 Reitz Lecture, (1999)

[21] Jae-jin Kim, Yang Dong Woo, An Empirical Study on the Influence of Start-ups' Technology-Commercialization Competence on Their Survival, Journal of The Korean Data Analysis Society, (2013), Vol.15, No.2, pp.1049-1061.
UCI: G704-000930.2013.15.2.042

[22] Jae-sik Lee, Chulung Lee, The Impact of Human Resource Competency on the Firm Performance of SMEs, Journal of Korean Institute of Intelligent Systems, (2016), Vol.26, No.6, pp.498-504.
DOI: https://dx.doi.org/10.5391/JKIIS.2016.26.6.498

