벤처창업연구 제16권 제3호 (통권75호) ISSN 1975-7557
Asia-Pacific Journal of Business Venturing and Entrepreneurship Vol.16 No.3 pp.73-86 e-ISSN 2671-9509

기계학습을 이용한 기업가적 혁신성 예측 모델에 관한 연구

정두희 (한동대학교 ICT창업학부 조교수)*
윤진섭 (한동대학교 경영경제학부 학부생)**
양성민 (한동대학교 AI Convergence & Entrepreneurship 학부생)***

국 문 요 약

이 연구의 목적은 기업가적 혁신성을 정확하게 예측하는 고도화된 분석 모델을 탐색하는 것이다. 기업가정신 연구 분야에서는 최초로, 데이터 과학적 접근방식에 해당되는 기계학습(Machine learning)을 이용해 기업가적 혁신성(entrepreneurial innovativeness)을 예측하는 모델을 제시한다. 예측모델을 구축하기 위하여 Global Entrepreneurship Monitor(GEM)의 62개국 22,099건 데이터를 이용한다. 27개 설명변수로 이뤄진 데이터 셋을 토대로 전통적 통계방법인 다중회귀분석과, 회귀트리, 랜덤포레스트, XG부스트, 인공신경망 등 기계학습을 이용한 예측모델을 구축하고 각 모델의 성능을 비교한다. 모델의 성능 평가를 위해 RMSE(Root mean square error), MAE(Mean absolute error)와 상관관계(Correlation) 등 지표를 사용한다. 분석 결과 5가지 기계학습 기반 모델은 모두 전통적 방법에 비해 우수한 성능을 보였으며, 예측 성능이 가장 좋은 모델은 XG부스트였다. XG부스트를 통한 기업가적 혁신성 예측에 있어서 기여도가 높은 변수는 창업가의 기회인지 및 시장 확장의 교차항 변수이며, 이는 신시장에서 기회를 획득하고자 하는 유형의 창업기업이 높은 혁신성을 보인다는 점을 확인했다. 이 연구는 고도화된 분석방법인 기계학습을 이용해 새로운 예측모델을 제시, 기업가정신 연구의 시야를 확장했다는 점에서 의의를 지닌다.

핵심주제어: 기계학습, 기업가적 혁신성, 예측모델, 데이터과학

I. 서론(Introduction)

수많은 연구들이 기업, 산업, 국가의 성장에 있어서 혁신이 중요한 동력이 된다는 점을 강조하며, 이러한 혁신의 원천이 무엇인지에 대한 탐구를 꾸준히 해왔다(Desyllas & Hughes, 2010; Koellinger, 2008; Yang et al, 2009). 기존 연구에서는 기업가의 경험적 속성(Cliff et al., 2006; Delmar & Shane, 2006), 사회인지적 요인(Ahlin et al., 2014), 조직적 요인(Urban, 2017), 환경적 요인(Mueller & Thomas, 2001) 등을 통해 기업 또는 기업가의 혁신성을 결정하는지를 밝히고자 노력을 해왔다. 기존의 연구들은 계량적 분석기법, 즉 로지스틱회귀(Koellinger, 2008; Farashah, 2015), 다중회귀분석(Soriano et al., 2012), 계층적(hierarchical) 회귀(Mueller, 2011), ANOVA(Nguyen, 2018) 등을 이용해 기업가적 혁신성을 형성하는 주요 변수들의 영향력을 분석해왔다. 이러한 접근은 연구자 디자인된 가설을 검증하는 방식으로 진행되는데, 연구자의 주관성 및 왜곡(bias), 인지 능력의 제한(Limitation of perception) 등 인간 능력의 한계로 인해 예측의 정확성을 극대화하는 데 제약이 생긴다는 점이 제기되고 있다(Prüfer & Prüfer, 2020). 이러한 전통적인 방법과 관련된 단점은 기업가정신 맥락에서 기업가적 혁신성 예측의 개선이 필요함을 시사한다.

한편, 지난 수십 년 동안 통계 및 컴퓨터 과학 분야에서 기계학습의 가능성이 검증되면서 예측 모델링 기술의 주목할만한 발전이 있었다(Putka et al., 2018). 기계학습은 정교한 알고리즘을 통해 데이터 속에 숨겨진 패턴을 파악하여 심층적 인사이트를 제공해주는 새로운 분석방법이다(Choudhury et al, 2021). 기업가정신 분야에서 기계학습 기법을 사용한 몇가지 연구가 있지만(Nasution et al., 2018; Tu et al., 2019; Tan, & Koh, 1996) 기업가적 혁신성 예측 연구를 고도화하기 위해서는 이러한 선진 방법의 활용에 대해 더 많은 이해를 할 필요가 있습니다. 기계학습 방법은 고도화된 예측 분석을 통해 기업가적 혁신성 예측에 있어서 새로운 가능성을 보여줄 것이다.

이 연구는 몇 가지 측면에서 중요한 기여점을 갖는다. 먼저, 이 연구는 기계학습에 기반한 분석 방법을 사용하여 기업가적 혁신성을 예측하는 모델을 제시한다. 모델의 정확도에 대한 진지한 고려 없이 가설 테스트 및 이론 구축에 초점을 맞추는 전통적 연구와 다르게, 주요 변수로 구성된 데이터에 기반하여 고도의 알고리즘을 통해 최고의 정확도를 제공하는

* 주저자, 한동대학교 ICT창업학부 조교수, profchung@handong.edu
** 공동저자, 한동대학교 경영경제학부 학부생, ezx0316@gmail.com
*** 공동저자, 한동대학교 AI Convergence & Entrepreneurship전공 학부생, didtjdals0708@gmail.com
· 투고일: 2021-05-10 · 수정일: 2021-06-11 · 게재확정일: 2021-06-23

예측 분석 모델을 개발하고자 한다. 이를 위해 전통적 분석방법으로 활용되는 다중선형회귀와 회귀트리, 랜덤포레스트 XG부스트, 신경망 등 4 가지 기계학습 방법을 사용하여 기업가적 혁신성 예측을 시도해본다. 이를 통해 전통적인 통계방법과 기계학습 모델의 성능을 비교해보고, 기업가적 혁신성을 예측하는 데 가장 성능이 좋은 모델을 확인하고자 한다. 또한 이 모델을 통한 예측에서 가장 기여를 많이 하는 중요 요소(변수)가 무엇인지를 식별하고자 한다.

이 연구는 단순히 새로운 시도의 소개에 그치는 것이 아니라 학자 및 실무자가 모델을 쉽게 복제하여 활용할 수 있도록 안내하는 데 중요한 역할이 있다고 판단한다. 이 연구에서 제시하는 모델은 기업가적 혁신성을 토대로 투자 의사결정을 내려야 하는 투자자와 국가 산업발전을 위해 창업자 지원 프로그램을 주관하는 정부기관 등이 참고하고 실제 상황에서 적용할 수 있도록 제시되기 때문에 실무적 기여도 또한 갖는다.

이 논문은 다음과 같이 구성된다. 첫번째 파트에서는 먼저 기업가적 혁신성의 중요성 및 결정 요인에 대한 문헌을 제시한다. 이후 새로운 예측 방법론으로서 기계학습의 개념 및 이 연구에서 사용할 기계학습 알고리즘에 대한 검토를 제시한다. 세 번째 파트에서는 알고리즘을 비교하고 혁신성 예측에 활용할 데이터 및 변수, 그리고 예측분석을 위한 방법론, 평가방법 등을 설명한다. 네 번째 파트에서는 분석 결과를 제시하고, 각 예측 모델의 성능을 분석, 가장 중요한 예측 변수를 식별한다. 다섯 번째 파트에서는 연구 결과에 대한 시사점 등을 논의한다.

II. 이론적 배경(Literature Review)

2.1 기업가적 혁신성 측정방식의 흐름 (Measuring entrepreneurial innovativeness)

많은 연구들이 기업가적 활동(Entrepreneurial activities)의 수행에 있어서 혁신성이 미치는 영향을 연구해왔다(Sciascia et al., 2013; Mueller & Thomas, 2001; Rezaei et al., 2012). 기업가(Entrepreneur)가 창의적인 아이디어에 기반하여 시장에 새로운 제품/서비스 출시 및 새로운 프로세스를 촉진하는 경향이 있을 때 기업가적 혁신성(Entrepreneurial innovativeness)을 갖는다고 일컬을 수 있다(Lumpkin & Dess, 1996). 기업가적 혁신성은 기업가에게 필요한 문제 해결 능력, R&D 추진력, 창조적 비즈니스 아이디어 생성을 탁월하게 수행해내는 능력과 밀접한 관련이 있다(Deniz & Godekmerdan, 2012, 김진영, 2019). 따라서 기업가적 혁신성은 창업 활동(Entrepreneurial activities)의 성공적 수행과 기업의 성장에 있어서 중요한 역할을 한다(Mueller & Thomas, 2001; Ünay & Zehir, 2012; 공혜원, 2018; 곽진만 외, 2017). Hurley & Hult(1998)은 기업가적 혁신성은 '기업 문화의 한 측면으로서 새로운 아이디어에 대한 개방적인 기업 문화를 형성하는 요인이라고 제시했다. 이러한 개방적 문화는 기존의 사고방식에서 벗어나 새로운 시장을 창조하기 위한 선결 조건이 된다(Savolainen, 2008). 이는 단지 새로운 창업기업의 초기 단계에만 국한되지 않고, 기존의 조직의 성공적 운영에 있어서도 유의한 역할을 한다(Antoncic & Hisrich, 2004; Lechner & Gudmundsson, 2014).

기존 연구들은 이렇게 기업의 성공적 운영 및 성장에 있어서 중요한 역할을 하는 혁신성을 예측하기 위한 다양한 시도를 해왔다. Koellinger(2008)은 로지스틱 회귀를 통해 2002~2004년의 Global entrepreneurship monitor 9,549명 데이터를 분석하여 기업가적 혁신성에 영향을 주는 요인을 탐색했다. 이 연구는 성별, 학력, 현재 고용상태, 자기효능감 등의 개인적 속성과 국가의 교육 제공 및 참여 수준(Educational attainment), 경제 개발 단계 등의 환경적 속성을 토대로 기업가적 혁신성의 출현을 예측할 수 있는 연구모형을 제시했다. 이 중 성별, 고용상태, 경제개발단계 등의 요인이 기업가적 혁신성을 형성하는 데 중요한 역할을 한다고 제시했다.

Vaillant & Lafuente(2019)은 1,984명의 카탈로니아 기업가 데이터를 통해 과거 기업가적 경험이 보고된 연쇄 기업가들의 후속 벤처의 혁신성에 미치는 영향을 조사한다. 헥맨(Heckman) 회귀를 통해 분석한 결과 과거의 기업가적 경험은 인지 스키마를 풍부하게 하여 기업가적 혁신성을 높인다는 점을 제시했다. Li et al(2018)은 SZEI에 141개, SYGVI에 29개의 대학원 기업에 소속된 학생 창업 기업인들을 대상으로 총 156건의 설문조사를 진행하여 얻은 데이터를 분석하여 대학원 기업가의 기업가적 혁신성에 영향을 주는 요인을 탐색했다. 계층 회귀 분석(Hierarchical regression analysis)로 분석한 결과 인적자본, 관리 자체 효율성, 과신, 기업가적 보상이 대학원 기업가의 기업가적 혁신성에 영향을 준다는 점을 발견했다.

기존의 기업가정신 연구들은 로지스틱회귀(Koellinger, 2008; Farashah, 2015), 다중회귀분석(Soriano et al., 2012), 계층적(Hierarchical) 회귀(Mueller, 2011), ANOVA(Nguyen, 2018) 등을 이용해 기업가적 혁신성 및 창업의도 등 기업가적 속성에 미치는 영향을 분석해왔다. 이러한 접근은 연구자가 사전에 가설로 설정한 변수간 관계가 통계적으로 유의한지 확인하는 방식으로 분석이 진행된다. 이러한 방식에는 몇 가지 한계점이 내포되어 있다.

먼저, 기존 연구에서는 모델의 정확성에 대한 극대화보다는 변수간 관계의 통계적 유의성에 초점 맞춘다. 독립변수와 종속변수 사이에 통계적으로 유의한 관계가 포착된다 하더라도, 종속변수를 예측하는 모델 자체의 정확도가 낮으면 변수간 관계가 갖는 의미는 퇴색된다. 모델의 정확도를 극대화하는 노력은 계량적 분석을 통한 이론의 개발에 있어서 근본적으로 수행해야 할 일이다.

또한, 기존 연구방법에서는 연구자가 지정한 가설에만 초점을 맞추어 분석을 진행하다 보니 연구자의 인식 한계가 모델의 한계로 귀결되는 문제점을 갖는다. Cyert & March(1963)는 인간의 역량 한계(Limited information processing capabilities)를

제시하면서 인간은 의사결정의 복잡성 및 불완전한 정보 등 때문에 덜 합리적 의사결정(Un-rational decision)을 하게 된다고 설명했다. 연구자는 시장에서 일어나는 모든 사안을 볼 수 없고, 그들의 시야(Field of vision)내로 인식(Perception)이 제한되며, 그 중에서도 현상적인(Phenomena) 것들만 인식(Perceive)되기 때문에 인식(Perceive)의 제한은 더 커진다. 즉, 연구자가 포착하지 못하면 현상 이면에 숨어 있는 중요한 관계를 가설로 제시하지 못하기 때문에 중요한 시사점을 간과할 가능성이 생긴다.

한편 최근에 데이터과학 분야에서 활용되는 기계학습에 다양한 분야의 학자들이 주목하기 시작했다. 기계학습의 접근은 방대한 데이터에 기반하여 복잡한 현상 속에서 유의한 패턴을 포착하고, 중요한 변수간 관계를 예측하는 데 활용된다(Mohassel & Zhang, 2017). 기계학습은 연구자의 설계에 의존하지 않고, 시스템이 심층적 분석 방법을 통해 데이터 속에 내포되어 있는 패턴을 발견해내기 때문에, 기존 통계방법보다 더욱 유연하고 심층적인 분석이 가능하며, 다양한 문제를 효과적으로 해결할 수 있는 잠재성을 갖고 있다. 하지만 기업가정신 및 혁신 분야의 기존 연구는 여전히 전통적인 계량분석 방법에 의존하고 있으며 기계학습의 활용은 부족한 실정이다. 이 연구에서는 기계학습을 통해 기업가적 혁신성을 보다 정확하게 예측할 수 있는 모델을 제시하고자 한다.

2.2. 기계학습(Machine learning)

인공 지능(Artificial Intelligence)는 인간의 학습능력, 추론능력, 인지능력 등을 컴퓨터 프로그램으로 구현한 시스템을 총칭한다. 기계학습은 인공지능의 중요한 구현 방법 중 하나이며, 1950년대에 기계학습으로 구현된 체커게임[1]을 개발한 Samuel(1959)은 기계학습에 대해 '기계학습은 명시적인 프로그래밍 없이 컴퓨터가 학습하는 능력을 갖추게 하는 연구 분야'라고 정의했다. 기계학습은 알고리즘에 기반해 데이터를 분석하고 스스로 패턴을 학습하며, 학습한 내용을 기반으로 판단이나 예측을 도출한다. 여기서 학습은 특정 사례에 대한 훈련을 통하여 축적한 경험을 바탕으로 컴퓨터 시스템의 성능을 개선하는 과정을 말한다(Jordan & Mitchell, 2015). 기계학습은 학자 또는 엔지니어가 컴퓨터에 필요한 모든 지식을 의도적으로 정의할 필요 없이 컴퓨터가 경험으로부터 지식을 스스로 수집하고 학습하여 성과를 창출해낸다(Goodfellow et al., 2016).

기계학습의 학습과정은 추상화와 일반화의 개념을 통해 수행된다. 추상화(Abstraction)는 데이터가 갖고 있는 정보가 의미 있는 개념으로 변환되도록 도와주는 지식표현(Knowledge representation)이다. 추상화를 통해 데이터 속에 내재된 의미 있는 특징들을 추출해내게 된다. 추상화된 지식을 유사한 상황에 적용할 수 있기 위해서는 일반화(Generalization) 과정이 필요하다. 일반화는 추상화될 수 있는 모든 개념 집합 중에서 중요한 발견을 제공하는 '관리 가능한 집합'으로 축소하는 과정이다. 일반화를 통해 모델은 실제 문제를 푸는 데 적용할 수 있는 형태로 갖추어지게 된다.

기존 사회과학 분야의 계량연구는 연역적 접근으로서 이론에 기반하여 연구자의 가정에 의한 가설개발 및 테스트 방식으로 이뤄지는데 반해, 기계학습은 이론을 기반으로 하기 보다는 귀납적 접근법으로서 데이터 학습에 의한 모델링을 통해 인사이트를 얻는다. 그렇다보니 기계학습을 적용하는 사회과학 연구에서는 이론에 대한 심도 있는 고찰보다는 알고리즘과 데이터 중심 모델링의 중요성이 큰 특징이 있다.

데이터 과학에 기반하는 기계학습 모델은 몇 가지 측면에서 회귀모델 등 기존 방법과 차별성을 갖는다. 첫째, 기계학습은 과거 지식을 배우고 개선하는 학습 메커니즘이 있다 보니 모델의 성과가 더욱 우수하다. 특히, 기계학습은 기존 방법론에 비하여 복잡한 비선형 관계(non-linear relationship)를 유연하게 모델링한다. 기존 선형모델의 경우 변수간 관계에 대하여 선형적 관계로 가정하고 분석을 하다보니 많은 입력값에 의존하는 함수를 찾아내거나 불규칙성 속에서 패턴을 찾아내는 문제에서 한계를 갖고 있다. 기계학습은 복잡한 데이터 속에서 비선형적인 관계를 잘 포착하여 더욱 심층적인 인사이트를 줄 수 있는 강력한 연구 도구다(Choudhury et al., 2021). 따라서 여러 연구들이 기존 모델에 비해 학습의 원리에 기반한 기계학습이 정확성, 오류 최소화 등 측면에서 우수한 성과를 냄을 제시하고 있다(Chulani et al., 1999; Dvir et al., 2006).

또한, 기존 모델과 달리 자가학습 및 자가 개선 기능이있어 입력과 출력 사이의 관계를 설정하는 데 훨씬 간단하고 효과적이다(Dvir et al., 2006; Wang & Gibson Jr, 2010). 기계학습을 이용하는 학자들은, 모든 가능한 입력에 대한 결과를 예상하여 그들이 직접 수동으로 프로그래밍하는 것보다, 원하는 입출력 동작의 예를 데이터로서 제시해줌으로써 시스템을 훈련시키는 것이 변수간 관계를 파악하는 데 있어서 훨씬 더 쉽고 유용하다는 점을 지속적으로 확인하고 있다.

이러한 장점을 지닌 기계학습은 시간에 따른 자동 개선, 불확실성 하에서 추론과 의사결정에 관련된 다양한 분야에서 활용되고 있다. 특히 컴퓨터 과학뿐만 아니라 인간 학습의 심리학적 연구, 적응 제어 이론, 교육 관행 연구, 조직 행동, 경제 등 복잡성이 큰 사회과학 영역에서도 유용하게 활용되기 시작했다. 기업가정신 분야에서는 아직 기계학습 접근을 활발히 도입하고 있지는 않지만 일부 소수의 연구들이 복잡한 문제를 해결하기 위하여 이 접근을 채택하고 있다.

Sabahi & Parast(2020)은 기계학습을 이용해 개인의 기업가적 지향과 기업가적 태도의 여러 측면을 토대로 프로젝트 성과를 예측하는 모델을 제안했다. 185개 관측치를 갖는 샘플을

1) 체커(Checkers): 체스판에 말을 놓고 움직여, 상대방의 말을 모두 따먹으면 이기는 게임이다. 체스와 다르게 체커 게임은 모든 말이 동일하게 대각선 앞으로만 한 칸을 전진할 수 있다.

통해 라소(Lasso), 능선(Ridge), 서포트 벡터 머신(SVM), 신경망, 랜덤포레스트 등의 기계학습 알고리즘을 사용하여 분석한 결과, 회귀 계열의 알고리즘인 라소 기반 모델이 프로젝트 성과를 가장 정확하게 예측하는 것으로 나타났다. 또한 자기효능감(Social self-efficacy), 비교성(Comparativeness) 및 적극성(Proactiveness) 등의 기업가적 태도 및 지향이 프로젝트 성과를 예측하는 데 중요한 요인임을 밝혔다.

Montebruno et al.(2020)은 영국의 기업가 역사 데이터를 이용해 기업가의 지위(entrepreneur status)를 예측하는 모델을 제시했다. 이 연구는 기존의 로지스틱 회귀를 10 개의 최적화된 기계학습 알고리즘인 최근 이웃(Nearest Neighbors), 서포트 벡터 머신(SVM), 가우시안 프로세스(Gaussian Process), 의사결정 트리, 랜덤 포레스트, 신경망 등과 비교했다. 분석 결과 최상의 정확도를 달성한 것은 신경망 계열의 딥러닝 RNN(Deep Learning Recurrent Neural Network) 모델이었다. 이 연구는 분류의 문제에 대해서 기계학습 알고리즘을 이용할 때 전통적으로 사용된 기술보다 더 나은 가치를 창출 할 수 있음을 보여준다.

한편, 기업가적 혁신성(Entrepreneurial innovativeness)을 예측하는 문제는 기업가정신 연구에서 중요한 비중을 차지함에도, 아직 기계학습을 채택한 연구는 존재하지 않는다. 기계학습의 우수한 패턴 인식 및 예측 능력은 방대한 데이터에 기반하여 기업가적 혁신성(Entrepreneur innovativeness)을 포함한 중요한 속성을 예측하는 도구로 활용될 수 있음에도, 아직 기업가정신 분야의 학자들에게는 커다란 관심을 받지 못하고 있다. 이 연구는 글로벌 기업가정신 데이터(Global entrepreneurship data)를 기반으로 기계학습 방법론을 이용하여 기업가적 혁신성을 예측하는 새로운 모델을 제시하고자 한다.

III. 분석모델

기계학습을 이용한 연구에서 분석모델을 선정하는 것은 중요한 절차다. 어떤 모델을 선택하느냐에 따라 정확도 성과가 달라질 수 있다. 하지만 기계학습 모델을 통해 데이터 분석 기반으로 이뤄지는 학습과 그에 따른 결과를 분석을 하기 전부터 확인하는 것은 일반적으로 어렵다. 따라서 데이터 종류 및 작업속성에 맞는 후보 모델을 선정하고 이들에 대한 분석 작업을 하면서 성과를 확인하면서 최적 모델을 결정 방식을 취한다(Heyburn et al., 2018). 기업가적 혁신성을 예측하는 이 연구의 경우 예측하고자 하는 종속변수가 연속형 변수이며, 이에 적합한 모델 후보 중에서 방식이 서로 중복되지 않으면서 우수한 성과를 내는 것으로 알려진 회귀트리, 랜덤포레스트, XG부스트, 인공신경망을 이 연구의 모델로 선정했다. 기계학습 예측 모델의 성과를 평가하기 위하여, 전통적인 방법인 선형회귀를 비교 모델로 설정한다.

3.1. 다중회귀분석(Multiple regression)

회귀분석의 개념은 프란시스 갈튼 경에 의해 1894년에 처음 제안되었다. 회귀분석은 연속형 변수들 사이의 관계를 나타내는 모형을 구한 뒤 적합도를 측정해 내는 분석 방법이다. 이는 수치 데이터를 분석하기 위한 가장 일반적인 방법이다. 단일종속변수에 대하여 단일 독립변수와의 관계를 분석할 경우 단순회귀(Simple regression)라고 하고, 여러 독립변수 사이의 관계를 분석할 경우는 다중회귀(Multiple regression analysis)라고 한다.

회귀분석은 시간에 따라 변화하는 데이터나 어떤 영향, 가설적 실험, 인과 관계의 모델링 등의 통계적 예측에 이용된다. 이 방법은 실제값과 예측값의 차이를 나타내는 오차를 최소화 하는 방향으로 변수들의 관계를 추정한다. 이는 변수가 지닌 특징과 결과 간의 관계에 대한 강도와 크기 추정치를 제공하기 때문에 결과에 대한 해석이 용이하다(Kumari & Yadav, 2018). 회귀분석을 통해 연구자는 한 모델에서 잠재적으로 중요한 모든 요인을 설명할 수 있다. 또한 수치형 관계를 필요로 하는 대부분 문제에 적용될 수 있으며, 결과를 빠르게 도출해낼 수 있는 장점을 갖고 있다(Marill, 2004). 한편, 회귀분석은 데이터가 정규분포를 따른다는 강한 가정에 기반하기 때문에 적용에 있어서 제약이 있다. 또한 수치 특징의 변수만 처리하기 때문에 범주형 데이터는 적용이 되지 못하고, 범주형 데이터를 수치형으로 변환 등 추가적 조정이 필요하다.

3.2 회귀트리(Regression Tree)

회귀트리는 수치형 변수의 예측을 위해 활용하는 의사결정트리(Decision tree)의 일종이며 기계학습의 대표적인 예측모델이다. 의사결정트리는 트리구조(Tree structure)를 활용하여 변수간 관계를 분석한다. 나무가 하나의 기둥에서 시작해 여러 갈래의 좁은 가지로 분화해가는 모습을 반영하기 때문에 트리라는 명칭을 얻게 되었다. 의사결정트리는 효율적인 재귀분할(Recursive partitioning)을 통해 분석이 진행된다(Torgo, 1997). 처음의 루트노드(Root node)는 전체 데이터셋을 의미하며 아무런 분할이 일어나지 않은 상태다. 여기서부터 알고리즘은 타깃이 되는 대상을 가장 잘 예측할 수 있는 특징을 선택하고 이 특징을 기준으로 데이터를 부분집합으로 분리한다. 각 부분집합은 종료기준에 도달할 때까지 계속해서 더 작은 부분집합으로 반복 분리를 한다. 이러한 분할은 부분집합의 데이터가 동질적이어서 잎(Leaf) 상태라고 판단되거나, 데이터를 구별할 특징이 남아있지 않거나, 미리 설정된 종료 조건을 충족할 때 종료된다.

의사결정트리는 분류 속성을 지닌 변수를 예측할 경우 분류트리(Classification tree), 수치형 변수를 예측할 경우 회귀트리(Regression tree)로 활용된다. 루트노드에서 재귀분할을 진행할

때 분할 후 결과의 동질성을 가장 크게 증가시키는 특징을 분할 조건으로 설정한다. 이 경우 분류트리는 엔트로피(Entropy)를 통해 동질성이 측정되는 반면, 회귀트리는 분산, 표준편차, 평균과의 절대편차 등의 통계량으로 동질성을 측정한다. 가장 일반적인 분할기준은 표준편차축소(Standard deviation reduction)이며 다음 식으로 정의된다.

이 모델은 의사결정트리의 빠르고 직관적이며 결과해석이 용이한 장점과 수치데이터를 모델링할 수 있다는 장점이 결합된 방법이다. 선형회귀처럼 연구자가 모델을 미리 설계하지 않고 컴퓨터가 자동 특징 선택을 통해 예측 작업을 수행한다. 반면 많은 양의 훈련 데이터가 준비될 때 우수한 성과를 낼 수 있으며, 트리 규모가 커지면 해석이 어려워지는 한계를 갖고 있다.

3.3 랜덤포레스트(Random Forest)

랜덤포레스트는 Breiman(1996, 2001)에 의해 제시되었으며, 다수의 의사결정나무(Decision tree)를 학습하는 앙상블(Ensemble) 종류다. 앙상블은 여러개의 서로 다른 예측모형을 생성한 후 모형들의 결과를 종합하여 하나의 최종 결과를 도출하는 방법을 의미한다. 이 방법론은 다수의 의사결정나무(Decision trees)에 기반하지만 배깅(Bagging) 방식을 통해 효율성과 정확도를 더욱 높인다. 배깅은 Bootstrap aggregation을 의미한다. 즉, 복원추출 방식인 부트스트랩(Bootstrap) 샘플링을 통해서 여러 훈련세트를 만들고, 각각의 훈련세트를 토대로 모델을 적용해서 그 결과값을 평균 냄으로써 분산을 줄이는 기법이다. 랜덤포레스트는 트리 간의 다양성을 증가시키는 방향으로 수행되는데, 두 가지 특징을 갖는다(Breiman, 2001). 첫째는 전체 데이터 세트의 부트스트랩 복제에 각 트리를 적합시킨다. 부트스트랩 복제는 교체로 가져온 동일한 길이의 사용 가능한 데이터 세트의 랜덤 하위 집합을 말한다. 각 샘플은 이전에 선택되 었는지 여부와 관계없이 전체 원본 데이터 세트에서 무작위로 선택되고 트리에 적합된다. 두 번째 방법은 각 노드에 대해 사용 가능한 전체 속성 중에서 작은 랜덤 속성의 부분 집합을 선택하고 이 부분 집합만 사용하여 최적의 분할을 검색한다. 이렇게 부트스트랩 적합과 속성의 하위 부분집합을 이용한 분할 등 두 가지 방식의 결합을 통해 랜덤포레스트는 예측 도구로서 효율적이고 우수한 성능의 결과를 만들어낸다(Efron & Tibshirani, 1985).

랜덤포레스트는 범주형 변수뿐만 아니라 연속형 변수 예측에도 적용되며 높은 예측력을 보여주기 때문에 널리 쓰이는 방법론이다(Lin et al., 2017). 특히, 무수히 많은 트리 구조를 갖고 있다 하더라도, 무작위 중복표본 추출(Bootstrapped)된 샘플이 서로 다른 특성변수의 조합으로 구성되어 있기 때문에 과적합과 다차원성의 오류에 빠지지 않는다. 반면 모델 안에서 어떤 과정으로 결과를 도출했는지 해석이 어렵다는 단점이 있다.

3.4 XG부스트(eXtream Gradient Boosting)

XG부스트는 부스팅(Boosting) 기법에 기반한 앙상블 모형이다. 부스팅은 여러 개의 약한 학습기(Weak learner)를 연결하여 강한 학습기(Strong learner)를 만드는 앙상블 기법이다(Freund & Shapire, 1997). 부스팅에서 리샘플링된 데이터셋은 상호보완적인 학습자를 생성하도록 구성된다. 각각의 새로운 모델은 이전 모델의 결함을 줄이는 방향으로 수정을 시도하며, 이렇게 순차적으로 모델을 구축하는 과정에서 성능을 높인다(Mitchell & Frank. 2017).

부스트 방식의 앙상블 기법은 Gradient boost, Light GBM, XGBoost 등이 있다. 이 중 XG부스트는 부스팅 알고리즘 중 그래디언트 부스팅 머신(Gradient Boosting Machine, GBM) 알고리즘을 병렬 학습이 지원되도록 구현한 알고리즘이다(Chen & Guestrin, 2016). GBM은 부스트 방식으로 최고 성능을 제공하는 지도학습 분야의 중요한 도구이지만 순차적으로 학습하도록 구현되어 있어 속도가 느리다는 점과 규제 기능이 없어 과적합이 쉽다는 문제를 가지고 있다(Mitchell & Frank, 2017). XG부스트는 GBM이 가진 속도 문제를 병렬 처리를 통해 극복하도록 구현되었으며, 과적합 규제 기능과 교차 검증 기능을 자체 내장하여 과적합 오류가 줄어든다. XG부스트는 분류, 회귀 및 순위 지정 작업에 모두 적용될 수 있는 확장성의 장점을 갖는다. 성능이 좋은 앙상블 또는 신경망 기법에 비해 속도가 빠르고 정확도가 높으며, 단일 시스템에서 수십억 개로 확장된 예제 분석이 가능하기 때문에 컴퓨터 자원활용률이 우수하다(Chen & Guestrin, 2016). 계산 과정을 추적하기가 어려워 결과에 대한 해석이 어려운 단점은 존재한다.

3.5 인공신경망(Artificial Neural Network)

인공신경망은 인간 뇌의 생물학적 뉴런(Neuron) 구조를 모방하여 만들어진 통계학적 학습 알고리즘이다(Kriegeskorte & Golan, 2019). 신경망은 정보를 처리 및 저장하는 뉴런(노드)와 정보를 전달하는 시냅스의 결합으로 이뤄진 네트워크이며, 뉴런이 학습을 통해 시냅스의 결합 세기를 변화시키는 방식으로 예측 또는 분류 등의 문제를 해결한다. 이러한 신경망 원리를 통계학적으로 구현한 것이 인공신경망 모델이다.

인공신경망에는 입력층(Input layer)과 출력층(Output layer), 그리고 은닉층(Hidden layer)이 존재한다. 은닉층(Hidden layer)이란 입력층(Input layer)과 출력층(Output layer) 사이에 존재하여, 현상 속에 내재된 패턴을 파악하기 위한 가상의 계산 단계라고 이해할 수 있다. 은닉층이 여러 층으로 구성된 것을 딥러닝(Deep learning)이라 하며 보다 심층적인 분석이 가능해진다. 입력층(Input layer)으로부터 은닉층(Hidden layer)으로 전달된 각 입력데이터에 대한 가중치(Weight)와 편차값(Bias)들이 은닉층(Hidden layer)에서 합산된다. 이렇게 합산된 값들은 활성화 함수를 거쳐 결과값으로 산출한다. 활성화 함수는 입

력 데이터와 각 입력데이터에 대한 가중치(Weight)를 계산한 출력값의 출력 여부를 결정한다.

인공신경망은 우수한 학습 능력, 일반화 능력, 그리고 적응 능력을 지니고 있으며, 특히 복잡하고 비정형적 데이터를 데이터 분석에서 탁월한 성능을 발휘한다. 데이터의 노이즈에 강한 내성을 갖고 있으며, 차원축소(Dimension reduction)에 획기적이고, 데이터 종류에 상관 없이 자가 구성할 수 있는 능력이 있어 다양한 어플리케이션에 쉽게 적용 될 수 있다(Hailesilassie, 2016).

인공신경망의 우수한 성능을 기대하기 위해서는 인공신경망을 학습시킬 수 있는 대규모 데이터 세트를 필요로 한다. 학습시킬 데이터가 부족하다면 인공신경망의 우수한 성능을 기대할 수 없다. 많은 연산량으로 모델 훈련 시간이 많이 걸리고, 하이퍼 파라미터가 많으며 결과 도출 과정에 대한 해석이 어려운 단점이 있다.

IV. 방법론

4.1 데이터 및 변수

이 연구에서는 기계학습 방법을 기반으로 기업가적 혁신성을 예측하는 모델을 목표로 한다. 기계학습 예측 모형의 학습을 위해서 R version 4.0.2를 사용하였다. Global Entrepreneurship Monitor(GEM) 데이터를 이용해 분석을 진행했다. GEM 연구는 설문조사를 통해 기업가정신의 특성, 국가 별 창업가 요인, 환경 요인, 창업기업의 특성에 대한 전반적인 데이터를 제공한다. 이 샘플은 GEM에서 2015년에 발간한 62개국 기업가정신 데이터 기반하며, 누락값을 제외하여 분석에 활용한 전체 샘플수는 총 22,099건이다.

데이터에 잡음이 많고 품질이 좋지 않는 데이터를 기계학습 모델에 사용하면 잘못된 결과를 도출할 수 있으므로 결측값을 처리하는 방법과 이상치를 제거하는 과정이 필요하다. 결측 값을 처리하는 방법을 크게 3가지로 나누어 볼 수 있는데, 결측 값이 나타나는 행을 처리하거나, 결측 값을 'Unknown'이나 -∞와 같은 라벨로 대체하는 방법, 결측 값의 비율이 상당한 변수들을 제거하는 방법이 있다. 이상치 제거는 변수의 분포에서 비정상적으로 벗어난 값을 제거하는 과정이다. '제3분위수 ± 1.5*사분위수 편차'를 기준으로 상/하위 극단을 확인한 뒤, 그 범위를 벗어나는 관측치를 제거하는 방식이다. 본 연구에서는 결측 값들을 제거하기 위해 각 sample에 있는 결측치를 확인 한 뒤, 결측치가 존재하는 행을 삭제하는 방법을 사용하였다. 또한 명목형 변수에 해당하는 가족구성원(hhsize)수의 이상치를 제거하였다.

이 연구에서 종속변수인 기업가적 혁신성 수준은 Koellinger(2008)의 연구가 제시한 방법을 차용했다.

Koellinger(2008)는 1997년에서 2005년까지의 GEM데이터를 사용하여 모방과 혁신의 분류 기준을 제시하여 기업가의 혁신성 수준을 계산하고, 혁신적인 기업가정신과 모방적인 기업가정신의 차이점을 제시했다. 이에 본 연구의 산출 변수인 기업가적 혁신성 수준은 고객의 제품 혁신성 인지 수준(TEACUST), 시장내 유사제품 판매 정도(TEACOMP), 경영방식 및 기술의 최신성(TEATECH)을 의미하는 변수들의 개별 점수 더한 후 평균을 구한 값으로 산출했다(표 1). 산출 식은 다음과 같다.

Innovativeness = (TEACUST + TEACOMP + TEATECH) / 3

<표 1> 종속변수 측정을 위한 세부항목

변수	타입	내용	응답 값	
TEACUST	명목형	고객이 해당 제품을 새롭고 참신하게 생각하는 정도	3 2 1	모든 고객이 새롭거나 참신하다고 생각 일부 고객이 새롭거나 참신하다고 생각 모든 고객이 새롭거나 참신하다고 생각 X
TEACOMP	명목형	시장 내 유사 제품의 판매 정도	3 2 1	시장 내에 유사제품 부재 일부 기업 유사제품 판매 많은 기업 유사제품 판매
TEATECH	명목형	경영 방식(procedure) 및 기술의 최신성	3 2 1	1년 내에 존재 1년에서 5년 사이에 존재 5년 이상

설명변수의 경우 크게 창업가 요인(Entrepreneur attributes), 창업기업 요인(Corporate attributes), 환경적 요인(Environmental attributes) 그리고 시장확장 요인(Market expansion attributes)으로 총 4가지 분류로 나뉜다. 세부적인 변수는 기존 연구에서 다루는 변수들을 참고하여 총 27개의 변수로 구성된다(Bogatyreva et al., 2019; Miralles et al., 2016; Autio et al., 2013; Zhao, 2005). 창업가 요인은 응답자 개인의 특징을 나타내는 요인이다. 먼저 인구통계학적 요인으로서 연령(age9c), 성별(gender), 가족구성원(hhsize), 교육 수준(GEMEDUC), 수입(GEMHHINC), 직업(GEMOCCU) 등을 포함했다. 연령(age9c)은 18세 이하부터 시작해서 65세이상까지 총 9개의 구간이다. 성별(gender)은 남성이 1, 여성이 0인 이항 변수로 표현되고 가족 구성원 수(hhsize)는 수치형 변수로서 가족 구성원의 분포를 나타내는 변수이다. 교육수준(GEMEDUC)은 응답자의 학력을 나타내는 변수이며, 이수하지 않음, 일부 중등교육 이수, 중등교육 이수, 고등교육 이수, 대학교육 이수로 나뉜다.

수입(GEMHHINC)은 응답자의 소득수준을 나타내는 변수로, 소득수준을 소득액을 기준으로 상/중/하 3분위로 구분하였다. 직업(GEMOCCU)을 나타내는 변수는 정규직, 비정규직, 실직 상태, 전업주부 등을 포함하여 총 7개의 범주로 나타낸 변수이다.

<표 2> 변수 설명

카테고리		변수	개요	유형
창업가 속성 (Entrepreneur attributes)	사회적 인지 속성 (Social cognitive attributes)	Fearfail	창업 실패에 대한 두려움	Binary
		Opport	향후 6개월간 당신이 살고 있는 지역에서 사업을 시작할 수 있는 좋은 기회가 있는가?	Binary
		Subskill	사업을 시작하는데 필요한 지식과 기술이 있다고 생각하는가?	Binary
	기업가적 경험 (Entrepreneurial experience)	Knowent	창업한 사람들을 개인적으로 알고 있는가?	Binary
		Discent	지난 1년간 창업활동을 한 경험이 있는가?	Binary
		Busang	지난 3년간 창업 기업에 자금을 지원해준 경험이 있는가?	Binary
	인구통계학적 속성 (Demographic attributes)	TEAOPTYP	창업 동기(1: 주도적 활동 ; 2: 수익 증대 3: 수익 유지 ; 4: 기타)	4 categories
		GEMEDUC	최종 학력(1: 미이수; 2: 일부 중등교육; 3: 중등교육; 4: 고등교육; 5: 대학교육)	5categories
		GEMHHINC	소득 수준 (1: Low 33%; 2: Mid 33%; 3: High 33%)	3 categories
		GEMOCCU	직업 유형 (1: 정규직; 2: 전업주부; 3: 실업; 4: 기타; 5: 계약직; 6: 자영업; 7: 학생)	7 categories
		Gender	성별 (남자 ; 여자)	Binary
		Hhsize	가족 구성원 수 (Mean:4.446; Median:4.000; s.d.3.377757; [min,max] : [0,8])	Integer
		age9c	나이 구간 (1:0~17; 2:18~24; 3:25~34; 4:35~44; 5:45~ 54; 6:55~64; 7:65+; 8: 결측 값)	8 categories
환경 속성 (Entrepreneurship environment)		CAT_GCR2	국가 발전 상태 (1: 요소주도형 경제 ;2: 효율주도형 경제; 3: 개혁 주도형 경제)	3 categories
		CULSUPyy	창업에 대한 사회의 가치 인식 (창업을 바람직한 경력으로 인식 + 성공한 창업가에 대해 높은 사회적 지위 인정 + 기업가정신 관련된 언론의 관심 등 항목들의 긍정적 응답 수)	4 categories
		Nbgoodc	창업을 바람직한 경력으로 인식하는가?	Binary
기업 속성 (Entrepreneurship Corporates)		TEAHITEC	기술 발전 수준(1: LOW- TECH; 2: MID-TECH;3:HIGH-TECH SECTOR)	3 categories
		TEASIC4C	회사 유형(1: 1차 산업; 2: 중화학 공업 ; 3:서비스 산업 ; 4:CONSUMER ORIENTED)	4 categories
		TEAEXP4C	해외시장 진출 지향성(1:76%-100%; 2:26% - 75%; 3:1 – 25 %; 4: None)	4 categories
		TEAOBEX	향후 5년내 고용에 대한 의지	Binary
		BABYBUSM	창업 후 생존 용이성: 최대 42개월 동안 회사를 운영하고 있는가?	Binary
시장 진출 (Market Entry)		TEAyyMEM_2	기술에 의존하지 않고 시장을 확장하는 방식	Binary
		MEM1XTec	첨단 기술 분야에서의 기존 시장 지향	Binary
		TEAyyMEM_3	신기술을 통한 시장 확장	Binary
		MEM1XOPP	창업가의 기회 인지 및 기존 시장 점유의 교차항	Binary
		MEMexp1XOPP	창업가의 기회 인지 및 시장확장의 교차항	Binary

사회적인지 요인도 포함했다. 기회인식(OPPORT)은 '앞으로 6개월 안에 당신이 살고 있는 지역에서 사업을 시작할 수 있는 좋은 기회가 있나' 에 대한 응답이며, 자기효능감(suskil)은 '새로운 비즈니스를 시작하는데 필요한 지식, 기술 및 경험을 보유하고 있는가' 에 대한 응답이다. 실패에 대한 두려움(fearfail)은 '실패에 대한 두려움이 창업에 방해가 되나'에 대한 응답, 인지된 네트워크(knowent)는 '당신은 지난 2년 동안 사업을 시작한 사람을 개인적으로 알고 있나' 에 대한 응답을 나타낸다. 해당 변수들은 부정이면 0 긍정이면 1을 뜻하는 이항변수이다. 기회유형(TEAOPTYP)은 창업을 시작하는 동기에 관한 변수로서, 주체적인 삶, 이익 증대, 현상 유지, 기타라는 응답을 가진 4가지 범주를 가진 변수이다. 이와 함께, 창업가 경험에 관한 변수도 포함했다.

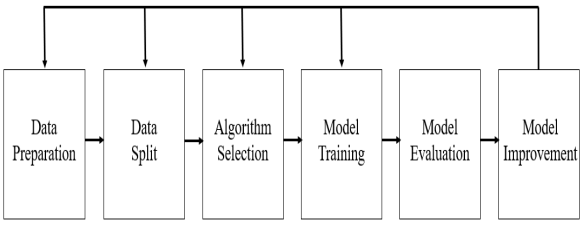
창업투자경험(busang)은 '지난 3년 동안 주식이나 뮤추얼 펀드의 구매를 제외한 다른 사람에 의해 시작된 새로운 사업을 위해 개인적으로 자금을 제공한 적이 있나'에 대한 응답으로 0이면 부정, 1이면 긍정의 의미를 뜻하는 이항변수이다. 창업 실패경험(discent)은 '지난 12개월 동안 당신이 소유하고 관리하는 사업, 어떤 형태의 자영업, 또는 상품이나 서비스를 누구에게든 팔거나, 문을 닫거나, 중단하거나, 그만두었는가?' 에 대한 응답으로, 0이면 부정, 1이면 긍정을 뜻하는 이항 변수이다.

창업기업 특성을 나타내는 변수들의 집합에는 기업의 기술수준(TEAHITEC), 회사형태(TEASIC4C), 시장확장방식(TEAyyMEM), 수출액 비중(TEAEXP4C), 기대고용(TEAOBEX)을 나타내는 변수들이 있다. 기업의 기술수준(TEAHITEC)는 노동력이 중요한 산업은 Low technology인 1, 중화학 공업과 같은 산업은 Medium-technology 인 2, 첨단 연구 및 기술을 요구하는 산업은 High technology인 3을 나타내는 범주형 변수이다. 회사형태(TEASIC4C)는 기업이 어떤 산업에 속하는지 나타내는 범주형 변수로 1은 농업과 광업과 같은 1차산업 형태, 2는 제조업 형태, 3은 B2B 형태, 4는 B2C형태의 산업을 뜻한다. 회사의 기대 고용(TEAJOBEX)을 나타내는 변수는 '5년간 5명 이상의 고용인원을 늘릴 계획이 있는지'에 대한 응답을 나타내는 변수이며, 회사 운영기간(BABYBUSM)을 나타내는 변수는 '최대 42개월까지 회사를 운영하는가'에 대한 응답을 나타낸다. 해당 변수들은 긍정의 경우 1, 부정의 경우 0의 값을 가지는 이항 변수이다.

환경 요인은 국가 발전 상태(CAT_GCR2), 문화적 지지도(CULSUPyy), 창업 선호도(nbgoodc) 등을 포함한다. 국가 발전 상태는 경제 개발 단계를 나타내는 변수로서, 총 3가지(Factor Driven, Efficiency Driven, Innovation Driven)로 분류되고, 창업 문화에 대한 지지도(CULSUPyy)는 해당 국가에서 창업에 대한 지지도가 어느정도 인지 나타내는 변수로서, 직업으로서의 창업가인식(NBGOODyy), 사회적 지위로서의 창업가인식(NBSTATyy), 창업에 대한 미디어의 인식(NBMEDIyy)과 같은 변수들에 대한 응답의 합을 나타내는 변수이다. 해당 변수는 응답한 수에 따라 0~3까지 값을 가지는 순서형 변수이다. 창업 선호도를 나타내는 변수는 '사람들은 창업이 바람직한 직업 선택이라고 생각하는가?' 라는 응답을 나타내며, 긍정의 경우 1, 부정의 경우 0의 값을 가지는 이항 변수이다.

시장 확장 방식(TEAyyMEM)은 시장을 확장하는 유형에 따라 기술에 의존하지 않고 시장을 확장하는 유형(TEAyyMEM2)과, 기술을 이용하여 확장하는 방식(TEAyyMEM3), 기술 기반 창업이지만 기존 시장 지향하는 유형(MEM1XTec)으로 구성되며, 각 변수는 긍정이면 1 부정이면 0을 뜻하는 이항 값을 갖는다. 이 연구에서는 시장확장 방식과 개인의 기회지향성 동기요인의 교차항을 변수에 포함했다. 창업가의 기회지향적 동기를 갖고 있으면서 시장확장을 추구하는 경우(MEMexp1XOPP)와 시장확장이 없는 기존 시장에서 기회지향적 창업을 추구하는 경우(MEM1XOPP)를 변수로 포함했고 이 두 변수는 각각 이항값을 갖는다.

기계학습에서 필수적으로 체크해야 할 사항은 과적합(Ovderfitting) 이슈다. 과적합은 기계학습의 모델 학습(Train) 단계에서, 학습 데이터(Train data)를 과하게 학습되어 실제 상황에 적용하기 어려워지는 경우를 말한다. 학습 데이타는 실제 데이타의 부분 집합이기 때문에 모델은 학습 데이타에 대해서는 오차가 감소하는 방향으로 학습되지만, 실제 데이타에 대해서는 오차가 증가하게 된다. 이러한 경우 모델의 일반화가 어려워지는 문제가 생긴다. 이를 방지하기 위한 방법은 데이터를 훈련(Train), 검증(Validate) 및 테스트(Test)의 세트로 분할하여 분석을 하는 것이다. 학습 데이터는 모델 학습 및 피팅에 사용되고, 검증 데이터는 파라미터 조정에 사용되며, 테스트 데이터는 예측 모형의 학습 과정에서는 배제하고 예측 모형이 학습된 이후, 예측 정확도 및 성능을 판단하는 용도로 사용된다. 전체 데이터는 임의로 발생시킨 난수를 사용하여 학습/검증 데이터와 테스트 데이터를 각각 7:3의 비율로 분할하였다. 이후 기술할 예측 모형의 성능 및 평가는 테스트 데이터를 대상으로 한다. 이 연구에서 기계학습 모델은 이러한 데이터 분할을 한 이후 각 알고리즘으로 모델 훈련(Training)을 진행한 다음 평가와 개선을 반복하여 최적의 모델을 가려내도록 했다. <그림 1>은 이러한 기계학습 모델링의 프레임워크를 나타낸다.



<그림 1> 기계학습 모델 구축을 위한 프레임워크

4.2 알고리즘 설정

이 연구는 회귀트리, 랜덤포레스트, XG부스트, 인공신경망 등 알고리즘을 사용하여 기계학습 기반 모델을 구축한다. 다중선형회귀 모델을 비교 모델로 설정한다. 기계학습의 중요한 준비단계 중 하나는 추정 및 하이퍼패러미터의 조정을 설정하는 것이다(Gu et al., 2018). 이를 위해 회귀트리는 R에서 제공하는 rpart 함수를 통해 모델링 하였다. 해당 모델의 패러미터는 anova 방법론으로 성정한다. 랜덤포레스트는 randomforest함수를 이용하여 구현하였고, 의사결정 나무 갯수인 ntree는 500개, 변수의 갯수를 뜻하는 mtry는 5로 설정하였다. XG부스트는 xgboost함수를 통해 구현하였으며, 최대 트리의 깊이인 max_depth는 6, 학습률을 나타내는 eta는 0.09, 모델 학습 횟수를 나타내는 nround는 400로 설정하였다. 인공신경망은 nnet함수를 이용하여 은닉층이 1개 이상인 다층퍼셉트론으로 구현되었다. 입력층은 27개의 설명변수, 은닉층은 5개의 은닉층으로 구성되어있다. 과적합 방지를 위해 학습률을 뜻하는 decay는 0.5로, 초기 가중치가 가지는 최대값을 05, 최대학습 횟수는 300으로 설정하였다. 또한 모델의 활성함수는 선형함수로 설정하였다.

4.3 평가 방법

기업가적 혁신성에 대한 예측 모델의 성능을 평가하기 위해 이 연구에서는 RMSE(Root mean square error), MAE(Mean absolute error)와 상관관계(Correlation)를 평가 기준으로 활용한다. 상관관계(Correlation)는 두 변수간의 관계의 유의성을 파악하는 지표로서, 예측값과 실제값과의 관련성을 확인하는 지표로 사용된다. 상관관계가 0.3 이상일 때 예측 성능이 충분하다고 평가될 수 있다(Wig et al., 2014). 평균 제곱근 편차(RMSE)는 모델의 실제 값과 예측 값 차이의 제곱 합을 데이터 전체수로 나눈 뒤 제곱근한 값으로, 예측 정확도 지표로 쓰인다. MAE는 예측값과 실제값의 절댓값의 평균을 의미하며 RMSE와 함께 수치 예측 모델의 지표로서 많이 사용된다(Botchkarev, 2018). 이와 같은 분석 프레임워크는 <그림 1>과 같이 도식화하여 표현할 수 있다.

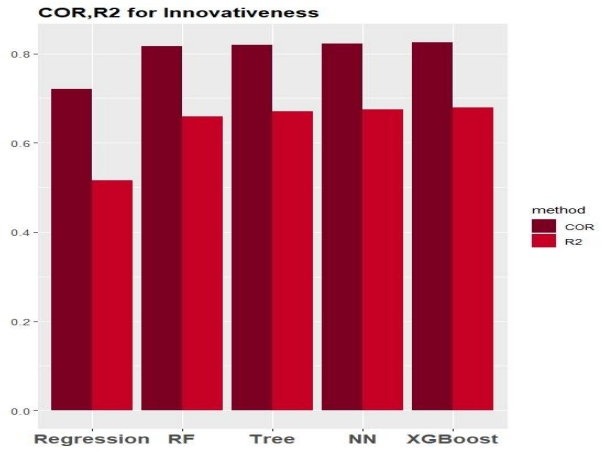
V. 결과

5.1 모델 성능

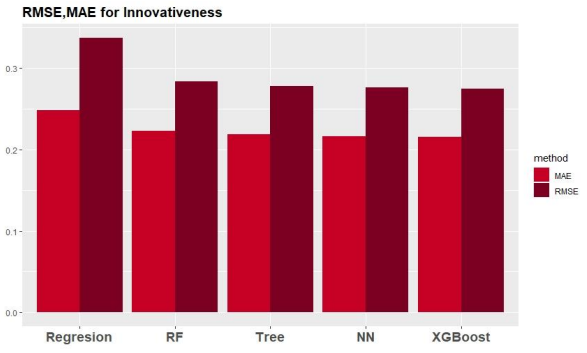
본 연구는 다중선형회귀, 회귀트리, 랜덤포레스트, XG부스트, 인공신경망 등의 모델을 이용해 기업가적 혁신성 예측모델을 구축했다. 기계학습 예측 모형의 학습을 위해서 R version 4.0.2를 사용했다. <그림 2>에서는 모델들의 성능 평가 결과를 나타내고 있다. 상관관계 지표는 XG부스트(0.825), 인공신경망 모형(0.823), 회귀트리(0.820), 랜덤포레스트(0.816), 선형회귀(0.720) 순으로 우수한 성능을 나타났다. R-squared 지표 또한 XG부스트(0.679), 인공신경망 모형(0.675), 회귀트리(0.671), 랜덤포레스트(0.659), 단순선형회귀모형(0.516)순으로 비슷한 결과를 보였다.

이와 함께, <그림 3>에서 볼 수 있듯이 RMSE 역시 XG부스트(0.274), 인공신경망 모형(0.276), 회귀트리(0.278), 랜덤포레스트(0.283), 선형회귀모형(0.337)순으로 우수한 결과를 보고했고, MAE 또한 XG부스트(0.216), 인공신경망 모형(0.216), 회귀트리(0.218), 랜덤포레스트(0.223), 선형회귀모형(0.248)순으로 비슷한 결과를 보여줬다. 즉, XG부스트 모형이 다른 모형들보다 실제 값과 오차가 적을 뿐만 아니라 예측 정확도 지표에서도 높은 성능을 보이고 있다. 이러한 결과는 기업가적 혁신성 예측에 있어서 기계학습 계열의 모델이 기존의 통계분석 방법보다 성능이 뚜렷하게 우수함을 나타낸다. 특히 앙상블 기법인 XG부스트 모형은 설명변수와 종속변수 사이의 복잡하고 계층적인 관계를 잘 포착하여 가장 우수한 예측 성능을 갖는 것으로 나타났다.

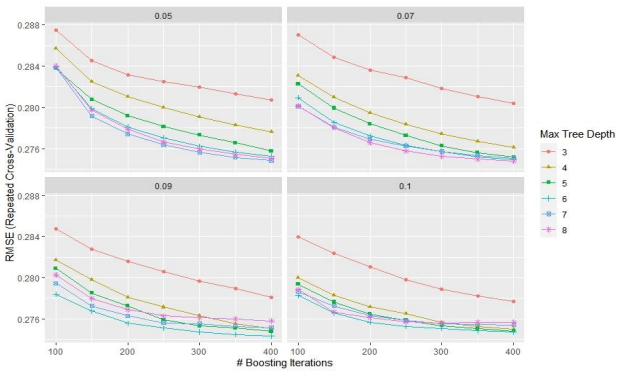
다음은 기계학습 모델 개선이다. 모델의 정확도를 높이기 위해서는 모델 성능을 최적화하는 파라미터 조건을 파악하는 작업이 필요하다. 일반적으로 처음부터 최적화 조건을 파악하는 것은 어렵다. 대신 하이퍼 파라미터들의 범위를 지정한 뒤, 교차 검증을 통해 모델의 최적화된 조건을 찾는 그리드탐색(Grid search)과정을 가질 필요가 있다. R에서는 이러한 작업을 수행하는 함수(GridSearchCV)를 제공한다. XG부스트 모델에 대한 그리드 탐색 결과는 <그림 4>와 같다. 이 결과에서 최대 트리의 깊이를 나타내는 max_depth는 6, 각 스텝마다 사용할 샘플의 비율을 나타내는 subsample은 0.63, 학습률을 나타내는 eta는 0.09, 모델의 iteration의 진행 횟수를 나타내는 nrounds는 400, 사용할 feature의 샘플 비율을 나타내는 colsample은 0.5인 경우, 예측 정확성이 가장 높았다.



<그림 2> 모형별 상관관계(Correlation), 결정계수(R-square)



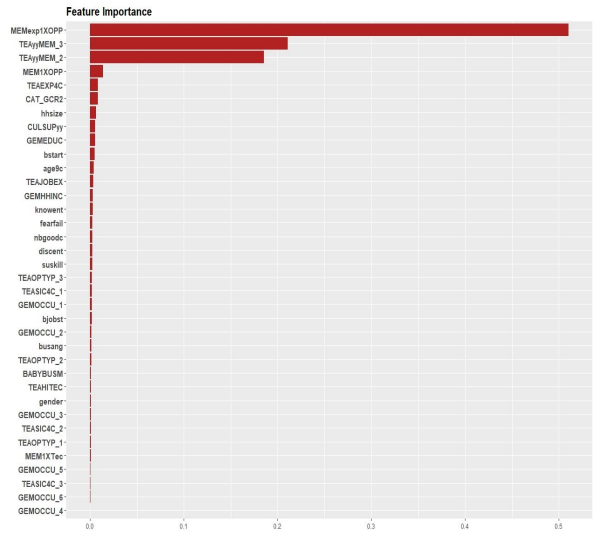
<그림 3> 예측 모형별 RMSE, MAE



<그림 4> 기업가적 혁신성 예측을 위한 XG부스트 성능 최적화

5.2 변수 중요도

이 연구의 또 다른 목적은 기업가적 혁신성을 예측하는 데 있어서 가장 기여도가 큰 변수를 확인하는 것이다. XG부스트의 경우 트리 모형에 각각의 변수를 포함할 때 얻게 되는 gain의 크기로 변수 중요도를 측정한다. 분석 결과, <그림 5>가 제시하는 것처럼 기업가적 혁신성에 가장 크게 기여하는 변수들은 '창업동기' 및 '시장확장 요인'과 관련이 많았다. 창업가의 기회인지 및 시장 확장의 교차항 변수(MEMexp1XOPP)가 혁신성에 대해 가장 높은 기여도(50.6%)를 보였다. 이는 신시장에서 기회를 획득하고자 하는 유형의 창업기업이 높은 혁신성을 보인다는 점을 나타낸다. 기존 시장에서 기회를 획득 추구 유형(MEM1XOPP)은 낮은 기여도(1.5%)를 보였다. 기술 기반으로 시장확장을 추구하는 경우(TEAyyMEM_3)가 두 번째(21.5%)였고, 기술에 의존하지 않고 시장확장 추구 유형(TEAyyMEM_2)은 이보다 낮은 수치로(18.2%) 그 뒤를 이었다. 매출액 대비 수출이 차지하는 비중 또한 혁신성 지수에 영향을 주었다. 산업적 시장확장뿐만 아니라 지리적 시장확장 역시 혁신성과 관련이 깊음을 나타낸다.



<그림 5> 기업가적 혁신성 예측을 위한 변수 중요도(XG부스트)

VI. 결론

본 연구는 글로벌 기업가 데이터 22,093건 데이터를 토대로 기계학습 방법론을 이용하여 기업가적 혁신성을 예측하는 모델을 구축했다. 창업가 개인의 요인과 창업 기업 요인, 환경적 요인 등 총 27개의 설명변수를 통해 기업가적 혁신성을 예측하는 모델들의 성능을 비교했다. 분석 결과, 전통적인 분석방법인 선형회귀에 비해 기계학습 모델인 회귀트리, 랜덤포레스트, XG부스트, 인공신경망 모델의 성능이 본 연구의 평가지표인 상관관계, RMSE, MAE 측면에서 우수하게 나타났다. 가장 우수한 성능을 보이는 모델은 XG부스트였다. 이러한 결과는 기업가적 혁신성을 예측하는 연구에 있어서, 기존의 선형회귀를 이용한 방식보다 기계학습 기반의 방식이 더욱 우수하며, 특히 XG부스트의 이용가치가 높음을 보여준다.

이 연구에서는 기업가적 혁신성을 예측하는데 기여를 많이 한 변수에 대한 탐색도 진행했다. 분석 결과 XG부스트를 통해 기업가적 혁신성을 예측하는 모델에서 창업가의 기회인지 및 시장 확장의 교차항 변수가 가장 높은 기여도를 갖는 것으로 나타났다. 즉, 새로운 시장에서 기회를 획득하고자 하는 창업자가 높은 혁신성을 갖게 됨을 보여준다. 또한 기술을 이용해 시장을 확장하는 창업가 유형 역시 높은 혁신성을 보였다. 이러한 결과는 기존 시장에서 기회를 찾는 것보다 새로운 시장을 창출/확장하는 기업가적 노력이 혁신성과 밀접한 관련을 가지며, 기술을 이용해 새로운 부가가치를 만들어내는 것이 높은 혁신성으로 연결됨을 보여준다. 기술에 의해 다양한 혁신이 일어나는 4차산업혁명 시대에 요구되는 혁신적 기업가의 조건을 이 연구는 구체적으로 제시한다.

본 연구는 다음과 같은 기여점을 지닌다. 첫째, 이 연구는 기업가적 혁신성(Entrepreneurial innovativeness)을 예측하는 문제에 대하여 기계학습 기술 기반 예측 모델이 지니는 가치를 기존 분석 방법과 비교하여 검증했다. 기계학습은 예측에 있어서 우수한 성능을 보유하고 있음에도 불구하고 그동안 기업가정신 연구에서는 적극적인 도입을 하지 않았다. 이 연구는 우수한 예측모델을 제시함으로써 기업가정신 분야 학자들이 기업가적 혁신성과 같은 주요 변수를 예측하는 연구를 수행하는 데 있어서 기존 방법론이 지닌 한계를 극복하며 보다 정교한 분석 모델을 사용할 수 있도록 한다. 이러한 결과는 예측연구의 방법론 측면에서 시야를 넓히는 데 기여한다. 이 연구에서 제시한 혁신성에 대한 예측모델은 기업가정신 활동의 성과를 평가하거나 창업기업의 가치를 분석하는 투자자 및 정부기관에게도 유용한 방법론으로 활용될 수 있을 것으로 기대한다.

이와 함께, 이 연구는 기계학습 방법을 통해 기업가적 혁신성을 예측하는 데 있어서 기여도가 가장 큰 변수를 탐색했다. 이 연구는 그동안 기업가정신 분야의 계량연구에서 제시한 변수들과 다른 결과를 보여주었다. 가령, Koellinger(2008)은 기업가의 혁신성을 높이는 요인으로 높은 학력과 높은 자신감(자기효능감)을 제시했으며, 개발도상국보다 선진국에서 혁신적 기업가 출현 가능성이 높다고 제시했다. 하지만 본 연구에서 사용된 XG부스트와 인공신경망을 통해 기업가적 혁신성에 가장 크게 기여하는 변수는 창업가의 기회인지 및 시장 확장의 교차항 변수였다. 앞으로의 기업가적 혁신성 연구에 있어서 해당 변수들을 고려한 심층적 분석이 필요함을 시사한다. 이 연구는 새로운 변수 관계를 발견하여 탐색연구로서의 의미 있는 역할을 했다고 볼 수 있다.

본 연구에는 한계점 또한 존재한다. 기업가적 혁신성에 대한 측정에 응답자의 자체 평가를 사용했다. 이는 주관적 평가

로 이뤄진 측정이다보니 객관적 측정 방법이 요구될 수 있다. 가령, 창업기업의 특허 등 지적재산권 등록 성과 등이 객관적으로 혁신성을 측정하는 대안이 될 수 있을 것이다. 이러한 객관적 측정을 통해 추가적 분석이 이뤄질 경우 보다 정확한 시사점을 제시할 수 있을 것으로 여긴다.

이와 함께, 이 연구는 2015년에 해당되는 GEM 설문을 바탕으로 모델 학습 및 평가를 진행했다. 이는 이 연구에서 모델에 포함하고자 하는 변수 중 해당 연도에만 존재하는 변수들이 있어서 (ex. 창업경험 등) 보다 넓은 범위의 데이터를 수집하지 못했다. 그러나 기업가적 활동은 기업가가 속한 환경적 맥락의 영향을 받는다. 다른 시간대의 데이터를 함께 분석할 수 있다면 모델 결과의 일반화가 보다 수월해질 수 있을 것이다.

이 연구는 귀납적 방식으로서 머신러닝 알고리즘을 이용하여 기업가적 혁신성에 대한 예측모델을 제시했다. 머신러닝 모델은 종속변수에 대한 예측 정보를 제공하지만 변수간 관계에 대한 정보는 충분히 제공하지 못하는 단점이 있다. 물론 부분의존도(Partial dependence plots, PDP)와 같이 모델의 예측 결과가 단일 설명 변수의 변화에 반응하여 어떻게 변하는 지 보여주는 기법을 통해 변수간 관계를 확인할 수도 있다. 추후 연구에서 이러한 기법을 이용해 기업가적 혁신성에 영향을 주는 변수들의 역학관계를 규명한다면 보다 구체적인 시사점을 이끌어낼 수 있을 것이다.

이 연구에서 기계학습 접근으로 시도한 새로운 방법이 기업가적 혁신성이라는 중요한 주제의 지평을 넓히는 데 기여하기를 바라며, 이 연구가 기업가정신 분야에서의 데이터 과학적 예측연구에 대한 관심을 확대하는 역할을 할 수 있기를 희망한다.

REFERENCE

김진영(2019). 기업가지향성이 중소기업의 기업가적 성과에 미치는 영향. *벤처창업연구*, 14(2), 83-93.

곽진만·양영석·김명숙(2017). 창업가 기업가정신 요인, 경영 관리적 요인, 자본적 요인이 창업성과에 미치는 영향 연구. *벤처창업연구*, 12(3), 119-133.

공혜원(2018). 창업경험 및 기업가정신 교육과 기업가 활동의도의 관계. *벤처창업연구*, 13(6), 129-141.

Ahlin, B., Drnovšek, M., & Hisrich, R. D.(2014). Entrepreneurs' creativity and firm innovation: the moderating role of entrepreneurial self-efficacy. *Small Business Economics*, 43(1), 101-117.

Antoncic, B., & Hisrich, R. D.(2004). Corporate entrepreneurship contingencies and organizational wealth creation. *Journal of management development*, 23, 518-550.

Autio, E., Pathak, S., & Wennberg, K.(2013). Consequences of cultural practices for entrepreneurial behaviors. *Journal of International Business Studies*, 44(4), 334-362.

Bogatyreva, K., Edelman, L. F., Manolova, T. S., Osiyevskyy, O., & Shirokova, G.(2019). When do entrepreneurial intentions lead to actions? The role of national culture. *Journal of Business Research*, 96, 309-321.

Botchkarev, A.(2018). Performance metrics(error measures) in machine learning regression, forecasting and prognostics: Properties and typology. *arXiv preprint arXiv*, 14, 45-79

Breiman, L.(2001). Random forests. *Machine learning*, 45(1), 5-32.

Breiman, L.(1996). Bagging predictors. *Machine learning*, 24(2), 123-140.

Chen, T., & Guestrin, C.(2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785-794.

Choudhury, P., Allen, R. T., & Endres, M. G.(2021). Machine learning for pattern discovery in management research. *Strategic Management Journal*, 42(1), 30-57.

Chulani, S., Boehm, B., & Steece, B.(1999). Bayesian analysis of empirical software engineering cost models. *IEEE Transactions on Software Engineering*, 25(4), 573-583.

Cliff, J. E., Jennings, P. D., & Greenwood, R.(2006). New to the game and questioning the rules: The experiences and beliefs of founders who start imitative versus innovative firms. *Journal of Business Venturing*, 21(5), 633-663.

Cyert, R. M., & March, J. G.(1963). *A behavioral theory of the firm*. Englewood Cliffs, NJ: Prentice Hall.

Delmar, F., & Shane, S.(2006). Does experience matter? The effect of founding team experience on the survival an sales of newly founded ventures. *Strategic Organization*, 4(3), 215-247.

Deniz, A., & Godekmerdan, L.(2012). Determining level of students' technological innovativeness: a case study. *Procedia-Social and Behavioral Sciences*, 47, 848-853.

Desyllas, P., & Hughes, A.(2010). Do high technology acquirers become more innovative?. *Research Policy*, 39(8), 1105-1121.

Dvir, D., Ben-David, A., Sadeh, A., & Shenhar, A. J.(2006). Critical managerial factors affecting defense projects success: A comparison between neural network and regression analysis. *Engineering Applications of Artificial Intelligence*, 19(5), 535-543.

Efron, B., & Tibshirani, R.(1985). The bootstrap method for assessing statistical accuracy. *Behaviormetrika*, 12(17), 1-35.

Farashah, A. D.(2015). The effects of demographic, cognitive and institutional factors on development of entrepreneurial intention: Toward a socio-cognitive model of entrepreneurial career. *Journal of International Entrepreneurship*, 13(4), 452-476.

Freund, Y., & Schapire, R. E.(1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1), 119-139.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y.(2016). *Deep learning*(Vol. 1, No. 2). Cambridge: MIT press.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, L., Wang, G., Cai, J., & Chen, T.(2018). Recent advances in convolutional neural networks. *Pattern Recognition,* 77, 354-377.

Hailesilassie, T.(2016). Rule extraction algorithm for deep neural networks: A review. *arXiv preprint arXiv*, 14(7), 371-381.

Heyburn, R., Bond, R. R., Black, M., Mulvenna, M., Wallace, J., Rankin, D., & Cleland, B.(2018). Machine learning using synthetic and real data: similarity of evaluation metrics for different healthcare datasets and for different algorithms. *In Data Science and Knowledge Engineering for Sensing Decision Support: Proceedings of the 13th International FLINS Conference (FLINS 2018)*, 11, 1281-1291.

Hurley, R. F. & Hult, T. M.(1998) Innovation, market orientation and organizational learning: an integration and empirical investigation. *Journal of Marketing,* 62(4), 42-54

Jordan, M. I., & Mitchell, T. M.(2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255-260.

Kim, J. Y.(2019). Impact of entrepreneurial orientation on small-and medium-sized enterprises' entrepreneurial performance: the mediating role of entrepreneurial knowledge position. *Asia-Pacific Journal of Business Venturing and Entrepreneurship*, 14(2), 83-93.

Koellinger, P.(2008). Why are some entrepreneurs more innovative than others?. *Small Business Economics*, 31(1), 21.

Kong, H. W.(2018). The Relationship between Entrepreneurial Experience and Entrepreneurship Education and Entrepreneurial Intention: Moderating Effect of Gender and Social Protection. *Asia-Pacific Journal of Business Venturing and Entrepreneurship*, 13(6), 129-141.

Kriegeskorte, N., & Golan, T.(2019). Neural network models and deep learning. *Current Biology*, 29(7), R231-R236.

Kumari, K., & Yadav, S.(2018). Linear regression analysis study. *Journal of the practice of Cardiovascular Sciences*, 4(1), 33.

Kwak, J. M., Yang, Y. S., & Kim, M. S.(2017). A Study on the Influence of Personal Characteristics, Business Management Factors, and Capital Factors on Entrepreneurial Performance: In the Center of Ameliorating Small Businesses Supporting Policy by Government in Beauty Service Industry. *Asia-Pacific Journal of Business Venturing and Entrepreneurship*, 12(3), 119-133.

Lechner, C., & Gudmundsson, S. V.(2014). Entrepreneurial orientation, firm strategy and small firm performance. *International Small Business Journal*, 32(1), 36-60.

Li, J., Qu, J., & Huang, Q.(2018). Why are some graduate entrepreneurs more innovative than others? The effect of human capital, psychological factor and entrepreneurial rewards on entrepreneurial innovativeness. *Entrepreneurship & Regional Development*, 30(5-6), 479-501.

Lin, W., Wu, Z., Lin, L., Wen, A., & Li, J.(2017). An ensemble random forest algorithm for insurance big data analysis. *Ieee access*, 5, 16568-16575.

Lumpkin, G. T., & Dess, G. G.(1996). Clarifying the entrepreneurial orientation construct and linking it to performance. *Academy of management Review*, 21(1), 135-172.

Marill, K. A.(2004). Advanced statistics: linear regression, part II: multiple linear regression. *Academic emergency medicine*, 11(1), 94-102.

Miralles, F., Giones, F., & Riverola, C.(2016). Evaluating the impact of prior experience in entrepreneurial intention. *International Entrepreneurship and Management Journal*, 12(3), 791-813.

Mitchell, R., & Frank, E.(2017). Accelerating the XGBoost algorithm using GPU computing. *PeerJ Computer Science*, 3, e127-e163.

Mohassel, P., & Zhang, Y.(2017). Secureml: A system for scalable privacy-preserving machine learning. In *2017 IEEE Symposium on Security and Privacy(SP)*, 19-38.

Montebruno, P., Bennett, R. J., Smith, H., & Van Lieshout, C.(2020). Machine learning classification of entrepreneurs in British historical census data. *Information Processing & Management*, 57(3), 102210-102257.

Mueller, S. L., & Thomas, A. S.(2001). Culture and entrepreneurial potential: A nine country study of locus of control and innovativeness. *Journal of business venturing*, 16(1), 51-75.

Mueller, S.(2011). Increasing entrepreneurial intention: effective entrepreneurship course characteristics. *International Journal of Entrepreneurship and Small Business*, 13(1), 55-74.

Nasution, M. D. T. P., Siahaan, A. P. U., Rossanty, Y., & Aryza, S.(2018). Entrepreneurship Intention Prediction using Decision Tree and Support Vector Machine. In *Proceedings of the Joint Workshop KO2PI and The 1st International Conference on Advance & Scientific Innovation*, 135-148.

Nguyen, C.(2018). Demographic factors, family background and prior self-employment on entrepreneurial intention-Vietnamese business students are different: why?. *Journal of Global Entrepreneurship Research*, 8(1), 1-17.

Prüfer, J., & Prüfer, P.(2020). Data science for entrepreneurship research: studying demand dynamics for entrepreneurial skills in the Netherlands. *Small Business Economics*, 55(3), 651-672.

Putka, D. J., Beatty, A. S., & Reeder, M. C.(2018). Modern prediction methods: New perspectives on a common problem. *Organizational Research Methods*, 21(3), 689-732.

Rezaei, J., Ortt, R., & Scholten, V.(2012). Measuring entrepreneurship: Expert-based vs. data-based methodologies. *Expert Systems with Applications*, 39(4), 4063-4074.

Sabahi, S., & Parast, M. M.(2020). The impact of entrepreneurship orientation on project performance: A machine learning approach. *International Journal of*

Production Economics, 226, 107621.

Samuel, A. L.(1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3), 210-229.

Savolainen, T.(2008). Organizational Trust and Leadership as Driving Forces for Innovativeness. In *Proceedings of the 13th International Conference on ISO9000 & TQM. Ho, S. (Ed.) Kuala Lumpur*, MY 71-72.

Sciascia, S., Clinton, E., Nason, R. S., James, A. E., & Rivera-Algarin, J. O.(2013). Family communication and innovativeness in family firms. *Family relations*, 62(3), 429-442.

Soriano, D. R., Guzmán-Alfonso, C., & Guzmán-Cuevas, J.(2012). Entrepreneurial intention models as applied to Latin America. *Journal of Organizational Change Management*, 25(5), 721-735.

Tan, S. S., & Koh, H. C.(1996). Modelling entrepreneurial inclination with an artificial neural network. *Journal of Small Business & Entrepreneurship*, 13(2), 14-24.

Torgo, L.(1997). Functional models for regression tree leaves. In *ICML,* 97, 385-393.

Tu, J., Lin, A., Chen, H., Li, Y., & Li, C.(2019). Predict the entrepreneurial intention of fresh graduate students based on an adaptive support vector machine framework. *Mathematical Problems in Engineering*, 8, 76841-76855.

Ünay, F. G., & Zehir, C.(2012). Innovation intelligence and entrepreneurship in the fashion industry. *Procedia-Social and Behavioral Sciences*, 41, 315-321.

Urban, B.(2017). Corporate entrepreneurship in South Africa: The role of organizational factors and entrepreneurial alertness in advancing innovativeness. *Journal of Developmental Entrepreneurship*, 22(03), 1750015.

Vaillant, Y., & Lafuente, E.(2019). The increased international propensity of serial entrepreneurs demonstrating ambidextrous strategic agility. *International Marketing Review*, 36(2), 239-259.

Wang, Y. R., & Gibson Jr, G. E.(2010). A study of preproject planning and project success using ANNs and regression models. *Automation in Construction*, 19(3), 341-346.

Wig, G. S., Laumann, T. O., Cohen, A. L., Power, J. D., Nelson, S. M., Glasser, M. F., Miezin, F. M., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E.(2014). Parcellating an individual subject's cortical and subcortical brain structures using snowball sampling of resting-state correlations. *Cerebral cortex*, 24(8), 2036-2054.

Yang, C. H., Motohashi, K., & Chen, J. R.(2009). Are new technology-based firms located on science parks really more innovative?: Evidence from Taiwan. *Research policy*, 38(1), 77-85.

Zhao, F.(2005). Exploring the synergy between entrepreneurship and innovation. *International Journal of Entrepreneurial Behavior & Research*, 11, 25-41.

Machine Learning for Predicting Entrepreneurial Innovativeness

Chung Doo Hee*
Yun Jin Seop**
Yang Sung Min***

Abstract

The primary purpose of this paper is to explore the advanced models that predict entrepreneurial innovativeness most accurately. For the first time in the field of entrepreneurship research, it presents a model that predicts entrepreneurial innovativeness based on machine learning corresponding to data scientific approaches. It uses 22,099 the Global Entrepreneurship Monitor (GEM) data from 62 countries to build predictive models. Based on the data set consisting of 27 explanatory variables, it builds predictive models that are traditional statistical methods such as multiple regression analysis and machine learning models such as regression tree, random forest, XG boost, and artificial neural networks. Then, it compares the performance of each model. It uses indicators such as root mean square error (RMSE), mean analysis error (MAE) and correlation to evaluate the performance of the model. The analysis of result is that all five machine learning models perform better than traditional methods, while the best predictive performance model was XG boost. In predicting it through XG boost, the variables with high contribution are entrepreneurial opportunities and cross-term variables of market expansion, which indicates that the type of entrepreneur who wants to acquire opportunities in new markets exhibits high innovativeness.

Keywords: Machine Learning, Entrepreneurial Innovativeness, Prediction Model, Data Science

* Assistant Professor, Handong Global University, School of Global Entrepreneurship and Information Communication Technology, profchung@handong.edu
** Handong Global University, Management & Econmics, ezx0316@gmail.com
*** Handong Global University, AI Convergence & Entrepreneurship, didtjdals0708@gmail.com