http://www.jsebs.org
ISSN: 2288-3908

The Journal of Society for e-Business Studies
Vol.25, No.4, November 2020, pp.77-109
https://doi.org/10.7838/jsebs.2020.25.4.077

Gradient Boosting 모형을 이용한 중소기업 R&D 지원금 결정요인 분석

Who Gets Government SME R&D Subsidy? Application of Gradient Boosting Model

강성원(Sung Won Kang)*, 강희찬(HeeChan Kang)**

초 록

본 논문에서는 그래디언트 부스팅 모형을 활용하여 정부의 중소기업 연구개발 지원 결정에 영향을 미치는 요인들을 파악하였다. 기존 연구가 사후적으로 정부의 연구개발 지원이 수혜기업에 미친 영향을 분석하는 것에 중점을 두었다면, 본 논문은 정부의 연구개발 지원 결정 방식을 파악하고, 그 방식이 기업에게 제공하는 유인을 분석하고자 하였다. 이를 위하여 본 논문은 지원금 결정에 영향을 미치는 다양한 잠재적 요인들을 선택하고, 기계학습 접근법을 활용하여 추정오차 축소효과가 큰 요인들을 선별하였다. 구체적으로 본 논문은 한국과학기술평가원이 구축한 국가연구개발조사분석 자료와 한국신용평가자료를 연결한 자료에 그래디언트 부스팅(Gradient Boosting) 모형을 적용하여 지원금 추정모형을 구축하였다. 본 논문에서 구축한 그래디언트 부스팅 모형은 선형회귀분석 응용모형에 비해 평균제곱근오차를 7.20% 축소할 수 있었다. 각 변수의 순열 중요도(permutation importance)를 분석한 결과 연구성과지표 및 연구개발비가 추정오차 축소에 기여가 큰 것으로 파악되었다. 그리고 각 변수의 부분의존도(Partial Dependence Plot: PDP) 및 SHAP 값(SHAP value: SHapley Additive exPlanation value)을 분석한 결과 연구성과지표가 좋고 연구개발비 지출이 큰 기업이 많은 연구개발 지원금을 받는 반면, 영업이익이 크고 자기자본회전율이 높은 기업은 적은 지원금을 받는 경향이 발견되었다. 본 연구의 결과는 현재 중소기업 연구개발 지원금 배분 방식이 연구성과지표 제고 및 연구개발투자 증가 유인은 제공하나, 기업 경영성과 제고 유인은 취약함을 시사한다.

ABSTRACT

In this paper, we build a gradient Boosting model to predict government SME R&D subsidy, select features of high importance, and measure the impact of each features to the predicted subsidy using PDP and SHAP value. Unlike previous empirical researches, we focus on the effect of the R&D subsidy distribution pattern to the incentive of the firms participating subsidy competition. We used the firm data constructed by KISTEP linking government R&D subsidy record with financial statements provided by NICE, and applied a Gradient Boosting model to predict R&D subsidy. We found that firms with higher R&D performance and larger R&D investment tend to have higher R&D subsidies, but firms

* First Author, Senior Researcher, Korea Environmental Institute(swkang@kei.re.kr)
** Corresponding Author, Associate Professor, Department of Economics, Incheon National University (henrykang@inu.ac.kr)
Received: 2020-09-15, Review completed: 2020-10-16, Accepted: 2020-10-29

with higher operation profit or total asset turnover rate tend to have lower R&D subsidies. Our results suggest that current government R&D subsidy distribution pattern provides incentive to improve R&D project performance, but not business performance.

키워드 : 연구개발, 중소기업, 기계학습, 그래디언트 부스팅, 정부보조금
R&D, SME, Machine Learning, Gradient Boosting, Government Subsidy

1. 서 론

본 논문은 기계학습 방법론인 그래디언트 부스팅(Gradient Boosting) 모형을 이용하여 어떤 요인이 정부의 중소기업 연구개발 지원 배분에 영향을 미치는지 파악하는 논문이다. 기존 관련 연구들을 보면 대부분 정부가 지원한 재원이 어떤 성과를 유발했는지, 즉 성과 중심으로 분석하고 있다. 그러나 본 논문에서는 정부의 연구개발 지원이 근본적으로 어떤 요인들에 의해 결정되는지를 파악하고자 한다. 기업들은 정부의 연구개발 지원을 더 효과적으로 받고자 적합한 자격요건을 갖추는 데 노력한다. 따라서 연구개발 지원 과정이 기업의 경영성과 제고와 정합적으로 이뤄진다면, 실제 지원금을 받은 기업뿐 아니라, 향후 수혜를 희망하는 관련 분야의 더 많은 기업들의 경영성과 제고를 유도할 수 있다.

일반적으로 정부의 연구개발 지원금 배분은 다양한 비시장적 요인들에 의해 복합적으로 이뤄진다. 따라서, 이들 요인을 선택하고 각 요인이 영향을 미치는 방식을 나타내는 함수형태를 선험적으로 선정하는 것은 분명 한계가 있다. 반면 기계학습법은 잠재적으로 영향을 미칠 수 있는 영향요인을 폭넓게 선정하고, 영향을 미치는 요인을 사후적으로 선별할 수 있다. 이러한 장점 때문에 기계학습법은 정부 지원의 결정요인을 분석하기에 적합하다고 볼 수 있다.

소규모 기업이나 고위험 기술개발의 경우 금융시장을 통한 자금 활용이 원활하지 않다. 따라서 정부는 연구개발 지원사업을 통해 중소기업의 이러한 문제를 해결해 준다. 그런데 정부는 지원 대상으로 재원조달이 어려운 기업을 우선 선정하는 경향이 있는데, 이 과정에서 경쟁력은 있지만 일시적으로 재원조달이 어려운 기업을 선정하기 보다는 단순히 경쟁력이 부족하여 재원조달이 만성적으로 어려운 기업을 선정할 위험이 있다. 이러한 방식으로 지원 대상 선정이 지속되면, 중소기업이 경영성과를 제고하여 재원부족을 스스로 해소할 유인을 약화시킬 위험이 있다. 이를 방지하기 위해서는 지원금 배분 방식이 중소기업이 경영성과를 제고할 유인을 제공하는 방식으로 이루어져야 한다. 그런데, 실제로 정부의 중소기업 지원금 지원 방식이 경영성과 제고 유인을 제공하는지 여부에 대한 실증연구는 아직 활발하지 않다.

지금까지 이러한 실증연구가 활발하지 못한 결정적인 이유는 정부의 재원 배분 방식이 “블랙박스”로 가려져 있어 불확실하기 때문이다. 따라서 기존의 미시경제학적 이론을 바탕으로 추정식을 도출하고, 적합한 변수를 선정하여 개별 변수의 영향을 추정하는 것으로는 정부의 지원금 결정요인을 모두 포괄할 수도 없을뿐더러, 이들 개별 요인이 지원금 결정방식에 어떻게 영향을 미치는지도 명확하게 파악하기 어렵다. 결과적으로 기존의 계량경제학적 접근은

추정모형의 과소적합(under-fitting) 문제를 유발하여 추정오차를 발생시킬 위험이 있다.

그런데 최근 사용이 활발해진 기계학습(Machine-learning) 방법론 중 '지도학습(supervised learning)'은 이러한 과소적합 문제를 해소할 가능성이 있다. 기계학습 방법론은 다양한 변인의 복합적인 영향을 추정에 반영하여 추정오차를 줄이고, 선험적인 변인의 선별을 최소화하기 때문에 다양한 변인을 파악할 수 있다. 이러한 기계학습은 개별 변인의 영향 파악이 어려운 약점이 있어서 정책적 시사점이 필요한 실증분석에는 광범위하게 사용되지 않았지만, 최근 기계학습 모형에서의 변수 영향 분석 기법이 빠르게 발달하면서 이를 해소할 수 있는 도구를 제공하고 있다. 본 연구에서는 부분의존도(PDP: Partial Dependence Plot)[7] 및 SHAP 값(SHAP value: SHapley Additive exPlanation value)[22, 25]을 이용하여 개별 변수 영향 분석이 어려웠던 기존 기계학습의 약점을 보완한다.

본 논문은 다양한 기계학습 방법론을 주어진 자료에 적용하여 지원금 규모를 추정하는 후보 모형군(群)을 선별하고, 이들 모형 중 추정오차가 가장 작은 추정모형을 선정한다. 그리고 선정된 모형에서 개별 변수의 추정오차 축소에 대한 공헌을 나타내는 순열중요도(permutation importance)[29]라는 기준으로 지원금 규모를 결정하는 중요 요인을 선정한다. 마지막으로 부분의존도 및 SHAP 값을 이용하여 개별 변수의 지원금 규모에 대한 영향을 정량적으로 분석한다. 기업의 경영성과와 관계되는 지표가 연구개발 지원금 결정에 큰 영향을 미치고, 경영성과가 좋은 기업에게 우선적으로 재원이 배분된다면 중소기업 연구개발 지원금 배분 방식은 기업경영 성과 제고 유인을 제공하고 있다고 할 수 있을 것이다.

본문의 구성은 다음과 같다. 제 2장은 중소기업 연구개발 지원금 성과평가에 관한 기존 실증연구를 정리하고 문제점을 파악한다. 제 3장에서는 본 논문에서 사용한 중소기업 연구개발 지원금 자료의 특성과 이 자료를 모형 분석에 사용하기 위해 전처리한 과정을 소개한다. 제 4장에서는 본 논문에서 사용한 그래디언트 부스팅 모형의 선정과정을 소개하고, 개별 변수의 순열 중요도, 부분의존도, SHAP 값 분석결과를 소개한다. 제 5장은 요약 및 결론이다.

2. 선행연구 검토와 문제 제기

기존의 기업 대상 정부 연구개발 지원 관련 실증연구는 지원을 받은 기업들의 성과에 대한 연구에 집중되어 있었고, 지원금 배분을 결정하는 요인에 대한 연구는 상대적으로 부족하였다. 이러한 기존의 연구는 투입된 정부 재원의 성과분석에는 소정의 성과를 거두었지만, 정부 지원금 배분 방식이 기업에 제공하는 유인에 대해서는 시사점을 제공하지 못하였다. 지원금 배분 방식에 대한 연구가 부족한 이유는, 선험적으로 변인을 선정하는 기존의 실증분석 방식으로는 정부지원 배분에 영향을 미치는 다양한 요인을 모두 포괄하기 어려웠기 때문이다. 최근 대두되고 있는 기계학습(Machine Learning) 방법론은 사전적으로 포괄적 요인을 분석에 포함할 수 있어서, 기존 실증분석 방식의 약점을 극복할 수 있다. 본 논문은 이러한 기계학습의 장점을 이용하여, 중소기업 연구개발 지원금 배분 요인을 파악하고자 한다.

정부의 연구개발 분야 기업지원 효과에 대한 연구는 오랜 역사에도 불구하고 아직 명확한 결론에 도달하지 못했다. 이론적으로는 정부의 연구개발 지원이 성공가능성이 낮은 기술에 대한 시장실패를 치유하여 부족한 자금을 공급하고 기업의 성과를 제고한다는 긍정론[27]과, 정부 지원으로 민간 투자가 구축되고, 결과적으로 민간의 생산성 제고 유인을 저해한다는 부정론[3, 21]이 공존한다. 이렇게 이론적인 논쟁이 지속되는 이유는 정부 지원의 이중적 성격 때문이다. 고급기술투자와 같은 고위험 자본시장에서는 투자자가 기업의 경쟁력을 판별할 정보가 부족하기 때문에, 경쟁력 있는 기업이 재원이 부족하여 경영성과를 제고하지 못하는 시장의 실패가 발생할 수 있다[9]. 이를 해소하기 위해서 정부가 재원을 제공하는데, 정부 역시 기업의 경쟁력을 판별할 정보가 부족하다[2]. 이 경우 정부는 재원 조달이 어려운 기업에게 우선적으로 지원금을 지급하는 경향이 있다[21]. 정부가 경영성과를 고려하지 않고 재원 부족만을 기준으로 지원금 지급을 결정할 경우, 경쟁력이 부족한 기업에게 지원금이 제공되어 성과가 부진한 문제가 우선 발생한다[3]. 뿐만 아니라 지원금 수급을 목적으로 경쟁하는 기업은 경영성과가 제고되어 재원 조달 여력이 강화되면 지원규모가 감소할 우려가 있어서, 경영성과를 제고할 유인이 약화되는 문제도 발생한다. 특히 중소기업 대상 지원정책은 재원 부족을 우선으로 지원대상을 선별하는 경향이 있어서, 두 가지 문제에 모두 취약하다.

이러한 이론적 논쟁들에 대한 실증연구는 지원 대상 선정 이후의 성과 분석에 집중되어 있고, 명확한 결론에 도달하지는 못하고 있다. 기존의 정부 연구개발 지원과 관련된 실증연구로는 정부지원금이 기업의 투자를 구축했는지 혹은 촉진했는지를 점검하는 연구가 있고, 한편으로는 지원금이 기업의 성과에 긍정적인 영향을 미쳤는지를 확인하는 연구가 있다. 정부 연구개발 지원으로 인한 민간투자 구축효과에 대한 실증연구의 결과는 확정적이지 않다[31]. 국내의 개별 중소기업 지원 정책이 혜택을 받은 기업의 성과에 미친 영향에 대해서는 다양한 주제로 실증연구가 진행되었는데, 세제혜택[13], R&D 지원[4, 5], 컨설팅 지원사업[19], 금융지원[3, 10], 수출입지원사업[28] 등의 영향에 대한 분석이 수행되었다. 이상의 연구에서도 긍정적인 영향 및 부정적인 영향을 뒷받침하는 결과들이 혼재한다.

이러한 연구들은 지원 대상 선정 이후의 사후적인 성과에 주목하기 때문에, 지원 대상 선정 과정에 대한 시사점을 제공하지는 않는다. 실제로 지원 대상 선정 방식이 기업에 제공하는 유인에 대한 실증연구는 찾아보기 어렵다. 최근의 근접한 연구로는 중소기업 지원 대상의 규모 제한이 고용에 미친 영향 분석[12]과 FGI 기법을 이용한 정부 지원자금 특징 분석에 근거한 중소기업정부 R&D 지원 전략 설계[17] 정도를 들 수 있다. 특히 중소기업 지원 대상 규모 제한의 파급효과에 대한 연구에서는 2013년 중소기업 지원범위 제한에 따라 해당 기업 고용증가율이 1.8~2.2% 하락하였다고 추정[12]하여 중소기업 지원방식이 기업의 고용 유인에 영향을 미치는 예를 보여주었다.

이렇게 지원 대상 선정 방식에 관한 실증연구를 찾아보기 어려운 이유는 정부 기업지원 방식에 대한 이론적인 기초가 확립되지 않아서, 추정식을 도출하기 어렵다는 한계가 있기 때문이다. 통상적인 경제학의 분석대상인 가계 혹은

기업과는 달리, 정부의 의사결정은 다양한 요인들의 복합적인 작용으로 이뤄진다. 그래서 가계나 기업과는 달리 정부의 의사결정에 대한 공리(Axiom)는 아직 학문적으로 정립되어 있지 않았다. 따라서 정부의 지원금 추정식을 사전적으로 결정할 미시적인 기초는 취약하다. 이러한 상황에서 제한된 이론적 기반을 이용하여 추정식을 도출할 경우, 정부지원금 추정에 필요한 요인을 누락하여 과소적합(under-fitting) 문제를 유발할 수 있다.

본 논문은 이러한 문제를 기계학습(Machine-Learning) 방법론 중 지도학습(supervised learning)을 사용하여 해소할 수 있음에 주목한다. 지도학습에서는 추정함수를 선험적으로 설계하지 않고, 주어진 자료를 이용하여 학습하는 방식을 채택할 수 있다. 구체적으로 지도학습 방법론에서는 종속변수에 영향을 미칠 수 있는 독립변수를 폭넓게 선택하고, 개별 변수와 종속변수 간의 관계도 자유도가 높은 함수를 선정한다. 그리고 추정오차를 최소화하는 함수의 모수 값을 파악하여 함수관계를 확정한다[8, 16]. 따라서 지도학습은 이론적인 기초가 명확하지 않을 때 과소적합의 문제를 완화하는 방법이 될 수 있다.

이러한 장점에도 불구하고 지금까지 지도학습은 정책연구에서는 폭넓게 활용되지 못하였다. 이는 지도학습에서 추정오차를 줄이기 위해 복잡한 모형을 사용하게 되면서, 개별 변수가 종속변수에 미치는 영향을 평가하기 어려워지는 약점이 존재하였기 때문이다. 그러나 최근 개별 변수가 종속변수에 미치는 영향을 평가하는 도구를 개발하여 기계학습 모형의 설명가능성을 제고하려는 연구가 활발하게 진행되고 있다[23, 24]. 본 논문은 이러한 도구 중 개별 변수의 독립적인 영향을 평가하는 부분의존도와 여타 변수와의 상호작용을 통해서 미치는 영향을 포괄하여 개별 변수의 영향을 평가하는 SHAP 값을 사용하여 변수의 영향을 평가하고, 그 결과로부터 정책적 시사점을 도출하고자 한다.

요약하자면, 본 논문은 다양한 요인을 포괄할 수 있는 기계학습의 장점을 이용하여 정부의 중소기업 연구개발 지원금 규모를 결정하는 요인을 실증적으로 파악하고자 한다. 구체적으로 본 논문은 다양한 지도학습 방법론을 적용하여 정부의 중소기업 연구개발 지원금을 추정하는 추정모형을 구축하고, 그 중 평균제곱근오차(RMSE)로 평가한 추정오차를 가장 축소할 수 있는 모형을 선정하여 정부의 연구개발 지원금을 추정한다. 이렇게 선정한 연구개발 지원금 추정모형에서 개별 독립변수의 연구개발 지원금 추정오차 축소 효과를 평가하여 추정 오차 축소 효과가 큰 중요 변수를 선정한다. 마지막으로 이렇게 선정한 중요 변수들이 지원금의 증감에 미친 영향의 크기와 방향성을 부분의존도 및 SHAP 값을 이용하여 분석한다. 특히 독립변수에 포함된 기업경영 성과 지표가 지원금 증감에 미친 영향을 파악하여, 정부 중소기업 연구개발 지원금 지원 방식이 기업 경영성과 제고 유인을 제공하고 있는지 여부를 진단한다.

3. 입력자료

3.1 원자료

본 논문에서 사용한 자료는 2015년 한국과학기술평가원에서 중소기업 연구개발 투자효과를 분석하기 위하여 구축한 자료이다[14]. 이 자료는 한국과학기술평가원에서 관리하는 국가연구

개발조사분석 자료와 한국신용정보(NICE) 자료와 연계하여 구축하였다. 국가연구개발조사 자료로부터는 2002~2010년간 정부로부터 연구개발 지원금이 지급된 연구과제 중 연구책임자가 중소기업에 소속되었던 지원금 자료를 수집하였고, 이를 기업을 단위로 재구성하였다. 그리고 한국신용정보(NICE)의 기업정보와 연계하여 기업경영지표를 부여하였다. 이 자료는 정부의 연구개발지원 실적과 기업의 경영성과 지표를 연계하여서, 연구개발 지원금과 기업의 특성을 함께 분석할 수 있게 설계한 장점을 지닌다. 이 자료는 7,538개 기업에 31,782건의 연구과제를 통해서 지원된 5.8조 원의 연구개발 지원금을 포괄한다. 이 자료는 내부자료로서 보다 최신 상황을 묘사한 유사한 자료는 입수가 어려웠고, 본 논문에는 사용하지 못하였다. 본 논문에서 사용한 연구개발 지원금 자료의 연도별 추이는 다음 <Table 1>과 같다.

<Table 1>에 수록된 표본에 포함된 변수는 167개이며 지원 연도, 기업 식별자료, 연구비 총액, 연구비 지급 건수, 한국신용평가 자료로부터 추출한 기업경영지표, 국가연구개발조사분석 자료로부터 추출한 연구성과지표로 구성된다. 연구성과지표에는 연도 별 연구성과 지표, 당해연도 연구성과지표, 1년 후 연구성과 지표, 1년 전 연구성과 지표, 그리고 연구성과지표 총계가 포함된다. 이들 중 지원금 지원 당시에 고려하기 어려운 당해연도 연구성과지표와 1년 후 연구성과 지표는 분석에서 제외하였다. 그리고 연도 별 연구성과 지표 역시 연구성과지표 총계와 정보가 중복되므로 제외하였다. 따라서 연구성과지표는 연구성과지표 총계와 1년 전 연구성과지표만을 포함하였다. 그리고 연구비 지급 건수는 타 변수의 영향과 관계없이 연구비 지급 액수의 대부분을 결정하는 현상이 발견되어서 분석에서 제외하였다. 이상의 원자료는 다음과 같은 전처리 과정을 거쳐서 기계학습의 입력자료로 전환되었다.

3.2 전처리

기계학습에서는 학습을 촉진하고 개별 변인이 추정 결과를 지배하는 현상을 방지하기 위해서 표본을 전처리한다. 본 논문에서는 1) 결측치가

〈Table 1〉 Government R&D for SME(2002~2010)

(Unit: Million Won, Count)

Year	Total R&D Investment	Gov R&D subsidy	Private R&D Investment	Number of Projects	Number of Firms
2002	302,284	194,029	108,255	1,487	923
2003	367,632	227,537	140,095	1,673	1,163
2004	484,693	318,159	166,534	2,486	1,485
2005	569,854	353,204	216,650	2,150	1,572
2006	626,824	483,638	143,186	3,129	2,106
2007	791,931	639,288	152,643	3,809	2,667
2008	1,049,246	837,206	212,040	5,208	3,420
2009	1,372,384	1,188,485	183,899	5,452	3,691
2010	1,791,765	1,569,508	222,257	6,383	4,615

많은 변수 및 관측치의 제거 2) 관측치가 소수의 값에 집중된 변수의 제거 3) 이상치의 제거 4) 결측치 보간 5) 표준화 5가지의 과정을 사용하였다. 통상적으로 기계학습에서 사용하는 변수는 왜도(Skewness)를 완화하고 차원을 낮추고 선형종속인 독립변수들을 제거하여 학습을 촉진하는 경향이 있으나[20] 본 논문에서는 이 세 가지 전처리는 수행하지 않았다. 본 논문에서 사용한 자료는 기업 자료로서 대부분의 변수가 왜도가 높은 성향이 있다. 이는 기업의 성과가 좋은 소수의 기업과 성과가 좋지 않은 다수의 기업이 공존하는 중소기업의 현황을 반영하는 현상으로 간주하였고, 따라서 왜도의 완화는 시도하지 않았다. 또한 본 논문에서는 주성분을 선험적으로 추출하지 않고, 모형 선택 과정에 주성분 추출을 포괄하는 모형을 포함하여 추정 오차 축소에 도움이 되는 경우에 주성분 분석을 사용하고자 하였다. 마지막으로, 본 논문에서는 선형 종속인 독립변수는 선험적으로 추출하지 않고 릿지 회귀분석 및 라소 회귀분석을 이용하여 예측오차를 최소화하는 독립변수를 선택하게 하였다.

우선 결측된 표본이 많은 변수와 결측된 변수가 많은 표본은 자료 구축 단계에서 오류가 있는 것으로 간주하여 분석대상에서 제외하였다. 구체적으로 20개의 변수는 결측치가 전체 표본의 55% 이상이어서 분석에서 제외하였다. 결측치가 55% 이상인 변수를 제외하고 부가가치 변수를 추가한 표본 중에서 68%의 표본은 결측치가 있는 변수가 전체 변수의 25.8% 이하였던 반면, 나머지 32%의 표본은 결측치가 있는 변수가 전체 변수의 36.9% 이상이었다. 따라서 결측치가 있는 변수가 전체 변수의 25.8%를 초과하는 표본은 분석에서 제외하였다.

이상치의 제거를 위해서 개별 독립 변수의 표본 내 1분위 값과 3분위 값 간의 구간의 일정 배수 이상이 되는 값을 포함하는 표본을 제외하였다. 표본을 보존하기 위해서 독립변수의 이상치 때문에 분석에서 제외하는 표본을 10% 이내로 제한하였고, 그 결과 개별 변수 중 하나 이상의 값이 표본 내 1분위 값과 3분위 값 간의 구간의 53배 구간 밖에 있는 값을 갖는 표본을 제외하였다.

관측치가 소수의 값에 집중된 변수는 상수와 유사한 분포를 가지며, 이러한 변수가 많이 포함되면 학습을 지연시키는 요인이 된다. 따라서 99% 이상의 표본에서 하나의 값을 갖는 4개의 변수는 상수로 간주하여 분석에서 제외하였다. 그리고 95%~98%의 표본에서 하나의 값을 갖는 변수가 5개 존재하였는데, 이 변수들은 모두 0 혹은 양의 값을 가졌다. 따라서 이 변수들은 0의 값을 가질 경우에는 0, 양의 값을 가질 경우에는 1의 값을 갖는 더미 변수로 전환하였다. 여타의 변수는 단일한 값을 갖는 표본이 89.2%를 넘지 않았고, 서로 다른 23개 이상의 값을 가지고 있어서 연속변수로 간주하였다.

결측치 보간(imputation)은 결측치가 발생한 기업 표본의 연간 추세를 우선 반영하고자 하였고, 기업 표본에서 연간 추세를 사용할 수 없는 경우에는 해당 변수의 연간 표본 중위값의 연간 추세를 반영하고자 하였다. 구체적으로 결측치가 있는 변수의 기업별 표본이 2개 이상의 관측치가 있는 경우에는 각 기업 표본의 가용한 관측치를 종속변수로 하고 상수와 연도변수를 독립변수로 하는 선형회귀분석을 수행하고, 그 회귀분석의 예측치로 결측치를 보간하였다. 이 경우 보간한 값이 이상치가 되어 분석에 영향을 미치는 위험을 억제하기 위해서 추정치가 개별 기업 표본 관측치의 최저값보다 작으면 개별 기업 표본의 최저값을 사용하였고, 추정치가 개별 기업 표본 관측이의 최대값보다 크면 개별 기업

표본의 최대값을 사용하였다. 그리고 기업별 표본의 관측치가 1개 이하일 경우에는 해당 변수의 연간 표본 중위값을 종속변수로 하고 상수와 연도변수를 독립변수로 하는 선형회귀분석을 수행하고, 그 회귀분석의 예측치를 사용하였다. 이 경우에도 보간한 값이 이상치가 되어 분석에 영향을 미치는 위험을 억제하기 위해서 보간한 값이 해당 연도 표본에서 해당 변수의 최대값을 초과하는 경우에는 해당 연도 표본에서 해당 변수의 최대값을 사용하였고, 보간한 값이 해당 연도 표본에서 해당 변수의 최소값보다 작은 경우에는 해당 연도 표본에서 해당 변수의 최소값을 사용하였다. 마지막으로 개별 기업 표본 내에서 특정 변수가 모두 결측치일 때는 해당 변수 연간 표본의 중위값을 사용하였다.

결측치 보간 후, 연속변수인 독립 변수는 평균을 제외하고 표준편차로 나누어 표준화하였다. 대안적인 방식으로는 표본 최대값에서 표본의 값을 뺀 값을 표본 최대값에서 표본 최소값을 뺀 값으로 나누어 변수의 값을 0과 1 사이의 값으로 치환하는 방식도 존재한다. 두 방식 모두 시도하였는데, 각각의 방식으로 표준화 한 변수들의 2변수 간 산포도와 표준화 이전 2변수 간 산포도가 유사하였다. 따라서 두 가지 표준화 방식에 큰 차이가 없어서 본 논문에서는 평균-표준편차를 이용한 표준화 방식을 사용하였다.

다만 의사결정나무 응용모형에서는 변수의 표준화가 추정의 성과에 영향을 주지 않는 경향이 있기 때문에[30], 본 논문에서 사용한 랜덤 포레스트 모형 및 그래디언트 부스팅 모형은 표준화를 하지 않은 입력자료를 사용한 모형의 추정오차와 표준화를 한 입력자료를 사용한 모형의 추정오차를 비교하여 추정오차가 작은 모형을 사용하였다. 본 논문에서는 두 모형 모두 표준화를 사용하지 않은 모형의 표준오차가 더 작았으므로, 표준화를 하지 않은 입력자료를 사용한 모형을 사용하였다. 따라서 본 논문에서는 그래디언트 부스팅 모형에서는 표준화를 하지 않은 입력자료를 사용하였고, 본 논문의 결과는 표준화 방식에 영향을 받지 않았다.

이렇게 구성한 표본 중 정부 연구개발 지원금 규모가 상위 2%인 표본은 추정오차를 축소하기 어려운 이상치로 파악되어서 분석에서 제외하였다. 이상치를 제외하지 않아도 분석결과는 유사하였다. 단, 이상치를 제외하지 않았을 경우 모형에 과소적합(under-fitting)현상이 발생하였을 가능성이 커서 이상치를 제외하였다. 이상의 전처리 과정을 거친 표본은 51개 변수의 연도-기업 관측치 40,688개로 구성된 표본이다. 이 표본이 기계학습에 사용되었다. 분석에 사용된 변수는 정부 연구개발 지원금, 연도, 기업경영지표 41개 변수, 연구성과지표 8개 변수이다. 이 중 정부 연구개발 지원금은 종속변수로, 나머지 50개 변수는 독립변수로 사용되었다. 분석에 사용된 변수의 목록과 표본통계량은 부록에 소개한다.

4. 분석결과

4.1 분석모형

지도학습(supervised learning)은 예측오차를 최소화하는 조건부평균 함수 $E(y|X, \theta, d)$의 모수값(θ)을 찾는 '학습'을 목적으로 한다. 예측오차를 정의하는 손실함수는 여러 가지 형태가 쓰이지만, 본 논문과 같이 연속변수인 종속변수를 추정하는 경우에는 평균제곱오차(Mean Squared Error) 및 평균제곱근오차(Root Mean Squared Error)를 가장 보편적으로 사용한다. 본 모형에서는 평균제곱근오차를 사용하였다. 전체 표본 중

80%는 학습표본(train set)으로 설정하여 학습에 사용하였고, 나머지 20%는 평가표본(test set)으로 설정하여 학습된 모형의 추정오차를 계산하여 모형의 성능을 평가하는 자료로 사용하였다.

지도학습 모형에서 사용하는 함수의 형태를 결정하는 하이퍼파라미터(Hyper-parameter)값(d)은 모수 추정 이전에 선택해야 한다. 본 논문에서는 하이퍼파라미터 선택을 위해서 교차검증(Cross Validation)을 사용하였다. 교차검증은 하이퍼파라미터 후보군 중 다음과 같이 계산한 추정오차의 기댓값이 가장 작은 하이퍼파리미터값을 선택한다. 우선 학습표본을 크기가 같은 여러 개의 부분 집합으로 나눈다. 그중 하나를 제외한 나머지 집단(학습표본: train set)을 이용하여 하이퍼파라미터 후보 값이 주어진 상태에서 오차를 최소화하는 모수를 찾는다. 그리고 이 모수를 활용하여 나머지 1개 부분집합(검증표본: validation set)의 추정오차를 구한다. 이러한 작업을 모든 가능한 조합에 대해 반복하여 각각의 조합에서 계산한 추정오차의 평균값을 구한다. 이렇게 구한 추정오차의 평균값이 가장 작은 하이퍼파라미터 후보값을 모형의 하이퍼파라미터 값으로 확정한다. 본 논문에서는 회귀분석 모형 및 주성분 분석 모형에서는 학습표본을 10개로 나누어 교차검증을 시행하였다. 그리고 계산부담이 큰 의사결정나무 기반 모형(랜덤포레스터 모형, 그래디언트 부스팅 모형) 및 신경망 모형에서는 학습표본을 3개로 나누어 교차검증을 시행하였다.

하이퍼파라미터 후보군을 선정하는 방식으로는 파라미터 공간 내의 가능한 모든 값을 사용하는 탐색 방식(Grid Search), 파라미터 공간 내에서 후보 값을 임의의 확률분포를 사용하여 추출하는 통계적 방식(Random Search), 파라미터 공간의 확률분포를 베이지안 방식으로 추정하고, 추정된 확률분포에 따라 후보값을 추출하는 베이지안 방식 세 가지가 쓰인다. 본 논문에서는 선형회귀분석 응용모형 및 주성분분석 응용모형 하이퍼파라미터 선정에는 탐색방식을, 하이퍼파라미터가 많은 의사결정나무(Decision-Tree) 기반 모형 하이퍼파라미터 선정에는 베이지안 방식 중 하나인 TPE(Tree Parzen Estimator)를 사용하였다[1]. 신경망 모형 하이퍼파라미터 중 완전 연결 은닉층(fully connected hidden layer)의 개수는 1개~5개 중에서 탐색방식을 이용하여 선정하였고, 은닉층의 개수가 주어진 신경망 모형의 하이퍼파라미터는 TPE를 사용하여 선정하였다.

지도학습 모형의 후보군(群)으로는 선형회귀분석을 응용한 2개 모형, 주성분분석(PCA: Principal Component Analysis)을 응용한 2개 모형, 의사결정나무(Decision-Tree)를 응용한 2개 모형, 그리고 신경망 모형 중 평가표본의 추정오차가 가장 작은 모형을 선택하였다. 본 논문에서 사용한 선형회귀분석 응용모형은 릿지 회귀분석(Ridge Regression)과 라소 회귀분석(Lasso Regression), 주성분분석을 응용한 2개 모형은 주성분 회귀분석(PCA: Principal Component Regression)와 부분최소자승법(PLS: Partial Least Squares regression), 의사결정나무를 응용한 2개 모형은 랜덤포레스트(RF: Random Forest) 모형과 그래디언트 부스팅(Gradient boosting) 모형, 신경망(Nueral Network) 모형은 모든 은닉층이 완전 연결 은닉층으로 구성된 모형이다. 개별 모형의 하이퍼파라미터를 교차검정을 통해 선택하고, 하이퍼타라미터가 확정된 모형의 평가표본에서의 평균제곱근오차를 비교하여 그 값이 가장 작은 모형을 추정모형으로 선택하였다.

이와 같이 다양한 조건부평균 함수로부터

적합한 모형을 추출한 이유는 종속변수와 독립변수 간의 관계의 비선형성에 따라서 적합한 모형이 다르기 때문이다. 선형회귀분석 응용모형은 종속변수가 독립변수의 선형결합으로 근사될 수 있음을 가정한다. 주성분분석 응용모형에서는 독립변수를 조합하여 추출한 새로운 독립변수(주성분)의 선형결합을 사용하면 종속변수에 근사한 추정치를 도출할 수 있음을 가정한다. 반면 의사결정나무 응용모형 및 신경망 모형은 독립변수와 종속변수간의 관계에 대한 선험적인 가정을 사용하지 않고, 비선형성이 존재할 경우에 이를 반영하여 종속변수를 추정할 수 있도록 복잡한 구조를 도입한다. 따라서 종속변수가 독립변수의 선형결합으로 근사하기 어려운 경우에는, 선형회귀분석 응용모형 및 주성분분석 응용모형을 사용하게 되면 비선형성을 반영하기 어려워서 과소적합(under-fitting) 문제가 발생한다. 반면 종속변수가 독립변수의 선형결합으로 근사될 수 있는 경우에는, 의사결정나무 기반 모형 및 신경망 모형을 사용하면 존재하지 않는 비선형성이 학습에 반영되어 추정에 사용하지 않은 평가표본에서 추정오차가 확대되는 과적합(over-fitting)의 문제가 발생한다. 본 모형에서는 비선형성을 선험적으로 평가하기 보다는, 다양한 수준의 비선형성을 반영할 수 있는 다수의 모형을 선정하고, 각 모형의 평가표본의 평균제곱근오차를 도출하여, 평균제곱근오차가 가장 작은 모형을 추정 모형으로 선택하는 방식을 사용한다. 아래에서는 추정모형의 후보군으로 활용된 개별 모형에 대해 간단하게 소개한다[16, 20].

4.1.1 선형회귀분석 응용모형

릿지 선형회귀분석과 라소 선형회귀분석은 선형회귀분석에 포함되는 독립변수의 조합 중 추정오차를 최소화하면서 과적합을 피하는 조합을 내생적으로 파악하는 방식이다. 릿지 회귀분석과 라소 회귀분석은 모든 독립변수를 추정에 포괄하여 계산하는 평균제곱오차와 추정계수의 값이 커지면 손실을 커지게 하는 벌칙항목을 더한 값을 목적함수로 설정하고, 이 목적함수를 최소화하는 변수 조합을 선별하여 추정에 사용한다. 릿지 회귀분석은 벌칙항목으로 추정계수의 제곱의 합을 사용하고, 라소 회귀분석은 벌칙항목으로 추정계수의 절대값의 합을 사용한다. 릿지 회귀분석 및 라소 회귀분석의 목적함수는 벌칙항목에 곱하여져서 추정계수의 값이 큰 변수에 대한 벌칙의 크기를 나타내는 모수를 하이퍼파라미터로 포함한다. 본 논문에서 이 모수를 10개 부분집합을 사용하는 교차검증을 실시하여 추정오차가 가장 작아지는 값을 탐색(Grid-search)하여 선택하였다. 릿지 회귀분석에 사용한 값은 0.04, 라소 회귀분석에서 사용한 값은 0이다. 이 값이 0일 경우 라소 회귀분석은 모든 독립변수를 사용하는 선형회귀분석과 동일하다.

4.1.2 주성분 분석 응용모형

주성분 분석 응용모형 중 주성분 회귀분석(PCR, Principal Components Regression)은 주성분 분석을 이용하여 독립변수의 차원을 최소화하고 이를 선형회귀분석에 사용하는 것이다. 주성분 회귀분석에서 독립변수 차원을 줄이는 것은 낮은 차원의 변수 조합으로도 독립변수의 분산을 설명할 수 있도록 주성분을 추출하는 것이며, 추정오차를 줄이는 것과는 무관하다. 반면 부분회귀분석(PLS, Partial least square)에서 독립변수 차원을 줄이는 것은 독립변수와 종속변수간의 공분산을 줄이기 위함이다. 이 두 분석에서 주성분의 개수가 하이퍼파라미터가

된다. 본 논문에서는 주성분의 개수를 10개 부분집합을 사용하는 교차검정을 실시하여 추정오차를 최소화하는 값을 탐색(Grid- Search)하였다. 즉 주성분이 1~50개인 주성분 회귀분석 모형 및 부분 회귀분석 모형을 교차검정하여 추정오차를 최소화하는 주성분의 개수를 도출하였다. 주성분 회귀분석에 사용된 주성분은 49개이며, 49개 주성분은 독립변수의 변동을 100% 설명하였다. 부분회귀분석에 사용된 주성분은 46개이고, 46개의 주성분은 독립변수의 변동을 99.84% 설명하였다.

4.1.3 의사결정나무 응용모형

의사결정나무 모형은 독립변수 공간을 계층적으로 나눈 후 분할한 공간 내의 종속변수의 대푯값을 종속변수의 추정치로 사용한다. 이 모형의 학습과정은 추정오차를 가장 작게 만드는 독립변수 공간을 결정하는 과정이다. 이러한 의사결정나무 모형은 학습표본에 과적합 되는 경향이 강하기 때문에, 여러 개의 소규모 의사결정나무를 만들어 각각의 의사결정나무의 추정치의 대푯값을 추정치로 사용하는 의사결정나무 응용모형이 주로 사용된다. 본 논문에서 사용한 모형은 랜덤포레스트 모형과 그래디언트 부스팅 모형이다.

랜덤포레스트 모형은 임의로 복수의 작은 의사결정나무를 만들어서 각각의 의사결정나무의 추정치의 평균값을 추정치로 사용한다. 랜덤포레스트 모형에서는 과적합을 피하고자 개별 의사결정나무는 여타 의사결정나무와 독립성을 유지하는 방식으로 생성한다. 그래디언트 부스팅 모형[7] 역시 복수의 의사결정나무를 만드는데, 개별 의사결정나무를 순차적으로 만든다는 점에서 랜덤포레스트 모형과는 차이가 있다. 하나의 의사결정나무가 도출되면 실측치와 예측치 간의 추정오차가 발생한다. 다음 단계에서의 의사결정나무는 이 추정오차를 종속변수로 사용한다. 이렇게 되면 두 번째 단계의 의사결정나무는 첫 번째 단계의 의사결정나무에서 발생하는 오차를 축소하는 역할을 하게 된다.

랜덤포레스트 모형과 그래디언트 부스팅 모형에서는 개별 의사결정나무의 수와 개별 의사결정나무의 특성에 대한 제약이 하이퍼파라미터가 된다. 본 논문에서는 랜덤포레스트 모형에서는 개별 의사결정나무에서 사용할 수 있는 독립변수의 개수의 상한 및 표본의 하한, 개별 분할 공간이 포함하는 표본의 하한, 의사결정나무의 수를 TPE(Tree Parzen Estimator) 방식을 이용하여 선택하였다. 그리고 그래디언트 부스팅 모형에서는 개별 의사결정나무에서 사용할 수 있는 독립변수의 수와 표본의 크기, 개별 의사결정나무의 층위의 상한, 의사결정나무의 수를 TPE 방식을 이용하여 선택하였다. 랜덤포레스트 모형에서는 개별 의사결정나무에서 사용할 수 있는 독립변수는 16개 이하, 개별 의사결정나무 구축에 사용할 수 있는 표본은 3개 이상, 개별 분할 공간에 포함되는 표본은 1개 이상으로 제한되었으며 개별 의사결정나무는 84개를 사용하였다. 그래디언트 부스팅 모형에서는 개별 의사결정나무에서 사용할 수 있는 독립변수의 수는 전체의 독립변수의 76.5%, 개별 의사결정나무에서 사용할 수 있는 표본은 전체 표본의 87.4%, 개별 의사결정나무의 계층은 9개로 제한되었고, 270개의 의사결정나무를 도출하여 사용하였다.

4.1.4 신경망 모형[11]

신경망 모형(Neural Network)은 다수의 노드(node)의 네트워크로 구성되는 층(layer)이 중첩되어 구성된다. 자료를 읽어 들이는 층을

입력층(Input layer), 최종 결과를 출력하는 층을 출력층(Output layer)이라고 하고 입력층과 출력층 사이의 층을 은닉층(hidden layer)이라고 한다. 은닉층의 각 노드는 이전 층의 노드의 출력값의 가중평균을 입력값으로 받아서 활성함수에 투입하여 출력값을 산출하고, 이 출력값과 같은 층의 노드들의 출력값의 가중평균이 다음 층의 노드의 입력값으로 사용된다. 본 논문에서 은닉층의 활성함수는 ReLu 함수를 사용하였다. 과적합의 문제를 해소하기 위해서 심층신경망 모형 분석에서 주로 사용하는 노드 중 일부를 임의 선택하여 학습에서 제외하는 드롭아웃(drop-out) 방식을 사용하였다. 그리고 계산 부담을 줄이기 위해서 가중치의 1회 개선에 다수의 입력 자료를 사용하는 배치(batch) 방식을 사용하였다.

본 논문에서 사용한 신경망 모형의 하이퍼파라미터는 학습의 회수, 배치에 포함되는 표본의 수, 은닉층의 개수, 은닉층을 구성하는 노드의 개수, 드롭아웃에서 학습을 시키지 않는 노드의 비중이다. 이 중 은닉층의 개수는 1개~5개를 임의로 선정하였고, 주어진 은닉층의 값에서 여타 하이퍼파라미터를 TPE(Tree Parzen Estimator) 방식을 이용하여 선택하였다. 이렇게 최적화된 5개의 신경망 모형의 평가표본에서의 평균제곱근오차를 비교하여, 그 값이 가장 작은 신경망 모형을 선정하였다. 선정 결과 은닉층이 1개인 신경망 모형의 평가표본에서의 표준제급곤오차가 가장 작았다. 이 모형은 학습을 600회 하였고, 각 학습에서 1회 가중치 개선에는 350의 표본을 사용하고 표본의 8.9%를 학습시키지 않는 방식으로 학습하였다. 1개 은닉층은 노드 55개로 구성되었다.

<Table 2>에서 볼 수 있듯이 은닉층의 수가 3개 이상이 되면 은닉층이 많을수록 평가표본에서의 평균제곱근오차는 증가하는 경향이 있었다. 특히 은닉층의 수가 5개가 되면 평가표본에서의 평균제곱근오차가 급증하여 뚜렷한 과적합 현상을 보였다.

4.2 분석결과

4.2.1 추정오차

4장에서 소개한 기계학습 모형을 평가표본에 적용하여 도출한 추정치와 실측치의 평균제곱오차(평가오차)는 <Table 3>과 같다. <Table 3>의 평가오차(Test Error)는 종속변수의 단위인 백만 원을 단위로 표기하였다. 의사결정모형을 응용한 2개 모형(랜덤포레스트 모형과 그래디언트 부스팅 모형)의 평가오차가 여타 모형의 평가오차보다 확연히 낮았다. <Table 3>의 세 번째 열은 각 모형의 평가오차와 릿지 회귀분석의 평가오차의 차이를 릿지 회귀분석의 평가오차로 나눈 값을 백분율로 표시한 값으로서, 각각의 모형이 릿지 회귀분석과 대비하여 평가오차를 개선한 정도를 나타낸다. 참고로 종속변수의 평가표본 표준편차는 155.057이었다. 따라서 그래디언트 부스팅 추정치는 표준편차보다 오차를 24.73% 개선하는 효과가 있었다. <Table 3>의

〈Table 2〉 Nueral Network: Root Mean Squared Error

(Uni: million won)

Nummber of Hidden layers	1	2	3	4	5
RMSE	119.101	120.712	120.000	122.057	148.170

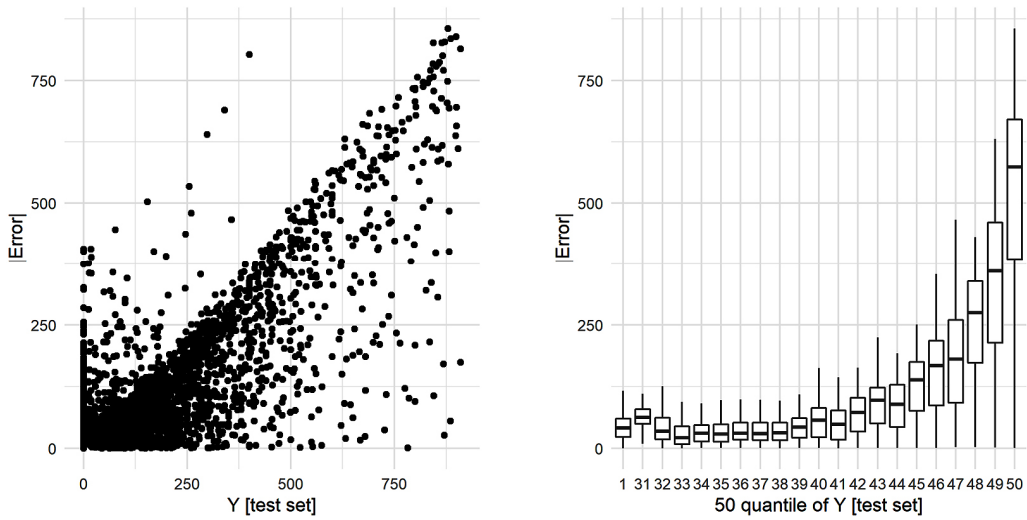
〈Table 3〉 Machine Learning Algorithm: Root Mean Squared Error

(Uni: million won, %)

Model	RMSE	RMSE Reduction .vs Ridge Regression
Gradient Boosting	116.713	7.20%
Random Forest	118.505	5.77%
Neural Network	119.101	5.30%
Partial Least Square	125.745	0.02%
Principle Component Regression	125.747	0.02%
Ridge Regression	125.767	0.00%
Lasso Regression	125.769	-0.002%

결과에 따라 본 논문에서는 정부 연구개발 지원금 추정 모형으로 그래디언트 부스팅 모형을 선정하였다.

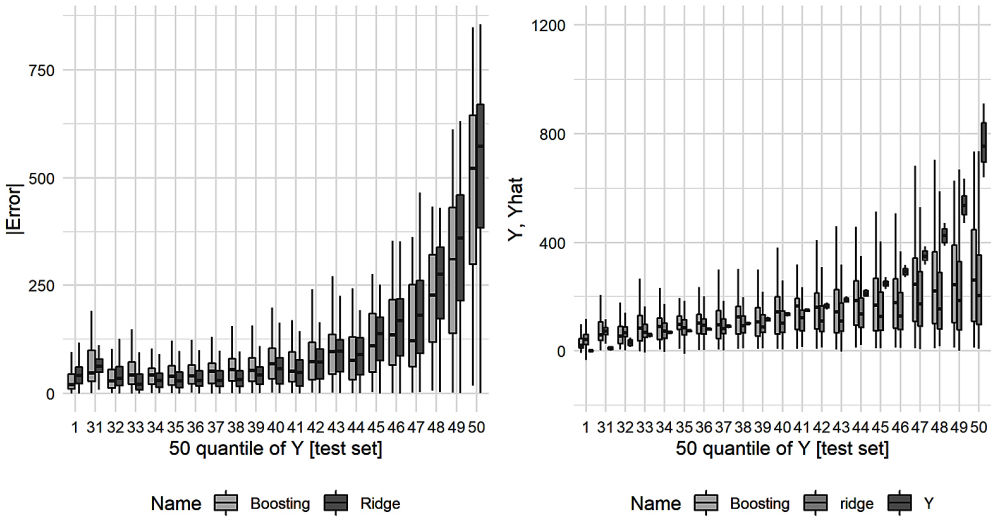
이처럼 의사결정나무 응용모형의 평가오차가 타 모형에 비해 낮은 이유는 <Figure 1>에서 볼 수 있는 바와 같이 종속변수인 중소기업 연구개발 지원금이 크기가 클수록 독립변수의 선형결합을 이용하는 추정모형의 오차가 증가하는 비선형성이 존재하기 때문이다. <Figure 1>의 왼쪽 그림은 선형회귀모형 응용모형 중 평균제곱근오차가 가장 작은 릿지 회귀분석(Ridge Regression) 모형의 추정오차의 절대값과 종속변수의 값 사이의 관계를 나타낸다. 그림에서 볼 수 있는 바와 같이 추정오차의 절대값은 종속변수의 값이 증가함에 따라 증가하는 성향이 확인된다. <Figure 1>의 오른쪽 그림은 종속변수를 100분위로 분할하여 각 분위 내의 추정오차의 절대값의 분포를 상자그림(boxplot)으로 나타낸 그림으로, <Figure 1>의 왼쪽그림에서 나타난 비선형성을 보다 간결하게 보여준다.



〈Figure 1〉 Absolute Value of Ridge Regression Prediction Error According to the Value of Dependent Variable: Scatter Plot and Boxplot Per 50 Quantile

<Figure 2>는 종속변수의 비선형성이 심한 구간에서 그래디언트 부스팅 모형의 추정치가 선형회귀분석 응용모형의 추정치보다 실측값과 상대적으로 가까운 경향을 보여준다. <Figure 2>의 왼쪽 그림은 평가표본의 종속변수를 50분위로 분할하고, 각 분위 내에서 릿지 선형회귀분석(Ridge) 추정오차의 절댓값의 분포 및 그래디언트 부스팅 모형(Boosting)의 추정오차의 절댓값 분포를 상자그림으로 나타낸 그림이다. 그리고 <Figure 1>의 오른쪽 그림은 평가표본의 종속변수를 역시 50분위로 분할하고, 각 분위 내 종속변수의 값(Y), 릿지 회귀분석 추정치(Ridge), 그래디언트 부스팅 모형(Boosting) 추정치의 분포를 상자그림으로 나타낸 그림이다. <Figure 2>의 왼쪽 그림은 <Figure 1>에서와 마찬가지로 두 모형 모두 종속변수가 커질수록 오차의 절댓값이 증가하는 추세가 존재함을 보여준다. 단, 릿지 회귀분석 추정오차에 비교해 보면 그래디언트 부스팅 모형 추정오차의 증가 추세가 약함을 확인할 수 있다. 그리고 <Figure 2>의 오른쪽 그림은 종속변수의 값이 큰 구간에서 그래디언트 부스팅 모형의 추정치가 릿지 회귀분석 추정치에 비해 상대적으로 커서 종속변수의 실측값과 가까운 경향이 있음을 보여준다.

의사결정나무 응용모형이 선형회귀분석 모형에 비해 성과가 좋은 원인은 개별 독립변수가 자체의 값에 따라 종속변수에 대한 영향력이 달라지는 비선형성이 존재하거나, 타 변수의 값에 따라 종속변수에 대한 영향력이 달라지는 변수 간 상호작용이 존재하기 때문이라고 할 수 있다. 선형회귀분석 모형에서 이와 같이 개별 변수의 비선형적 영향이나 상호작용이 존재할 경우에는 개별 독립변수를 비선형 함수로 변형하여 포함시키거나, 독립변수의 곱과 같이 독립변수 간 상호작용을 나타내는 변수를 회귀식에 명시적으로 포함하여야 한다. 그러나 의사결정나무 응용모형에서는 별도의 변수를 포함하지 않아도 비선형성 및 상호작용을 반영할 수 있다.



〈Figure 2〉 Distribution of Absolute Value of Prediction Error, Predicted Value, and Actual Value of Dependent Variable Per 50 Quantile of Dependent Variable: Gradient Boosting vs. Ridge Regression

이와 같은 추정오차 개선효과는 변화 폭이 크고 비선형성이 강한 종속변수를 기계학습을 이용하여 추정한 선행연구에서도 반복적으로 관찰되었다[15].

4.2.2 변수 중요도

본 논문에서는 그래디언트 부스팅 모형은 개별 독립변수가 추정오차를 축소하는데 기여한 여부를 순열 중요도(permutation importance)를 이용하여 평가한다[6]. 순열 중요도는 특정 독립변수의 배열을 임의로 뒤섞어서 그 독립변수의 정보를 제거하고 모형을 추정하여 얻은 추정오차와 모든 독립변수의 실제 값을 이용하여 얻은 추정오차의 격차를 의미한다. 전통적으로 그래디언트 부스팅 모형과 같은 의사결정나무 응용모형을 사용한 분석에서는, 의사결정나무 공간 분할을 기준으로 개별 독립변수의 중요도를 평가하였다. 구체적으로 의사결정나무 기반 변수 중요도는 의사결정나무에서 공간 분할 기준으로 활용된 독립변수가 추정오차를 개선하는데 기여한 정도를 그 독립변수로 분할되기 이전 공간 내에서 종속변수의 분산과 분할된 이후 공간 내 종속변수의 분산의 차이로 평가하였다. 그런데 표본 내 변수의 값이 많은 연속변수는 공간 분할에 보다 빈번히 사용되는 경향이 있기 때문에, 이러한 방식은 연속변수의 중요도를 과대평가하는 경향이 있다[29]. 특히 본 논문에서 사용한 독립변수 중 5개의 변수가 0과 1의 값을 갖는 이항변수이므로, 변수 간 중요도 비교에는 순열 중요도가 보다 적합하다.

본 논문에서 사용한 그래디언트 부스팅 모형의 독립변수는 50개이다. 이들 중 연도를 제외한 49개 변수의 순열 중요도를 도출하여 4개의 군집으로 구분하였다. 클러스터의 수를 바꾸어 클러스터를 나누어도 가장 중요도가 높은 3개 집단의 평균은 안정적으로 유지되었다. 군집의 기준은 평균거리를 사용하였다. 각 군집의 특징은 <Table 4>와 같다. 그리고 최고(最高)중요도군(Very High), 고(高)중요도군(High), 중(中)중요도군(Middle)에 포함된 변수는 <Table 5>와 같다. 상위 3개 군집에 포함된 변수는 10개에 불과했지만, 이 10개의 변수를 제외하면 개별 변수의 순열 중요도가 0.4에 미치지 못했다. 실제로 저(低)중요도군(Low) 중에서 순열중요도가 가장 높은 변수는 해외특허총합계 변수(op_total)였는데, 이 변수의 순열중요도는 0.393이었다. 이는 릿지 회귀모형 대비 그래디언트 부스팅 모형의 평균제곱근오차 개선효과 9.054의 4.3%에 불과한 값이다.

〈Table 4〉 Clusters of Importance of Features in Random Forest Model

Cluster	Average Permutation Importance	Count
Very High	19.767	1
High	7.334	3
Middle	0.805	6
Low	0.053	39

<Table 5>의 결과를 보면 연구성과지표가 정부 중소기업 연구개발 지원금 추정의 오차를 낮추는 데 기여도가 큰 것으로 파악된다. 분석에 포함된 총 8개 연구성과지표 6개가 최고(最高)중요도군(Very High), 고(高)중요도군(High), 중(中)중요도군(Middle)을 구성하는 10개 변수에 포함되었다. 반면 분석에 포함된 총 기업경영지표 41개 중 4개만이 최고(最高)중요도군(Very High), 고(高)중요도군(High), 중(中)중요도군(Middle)에 포함되었다.

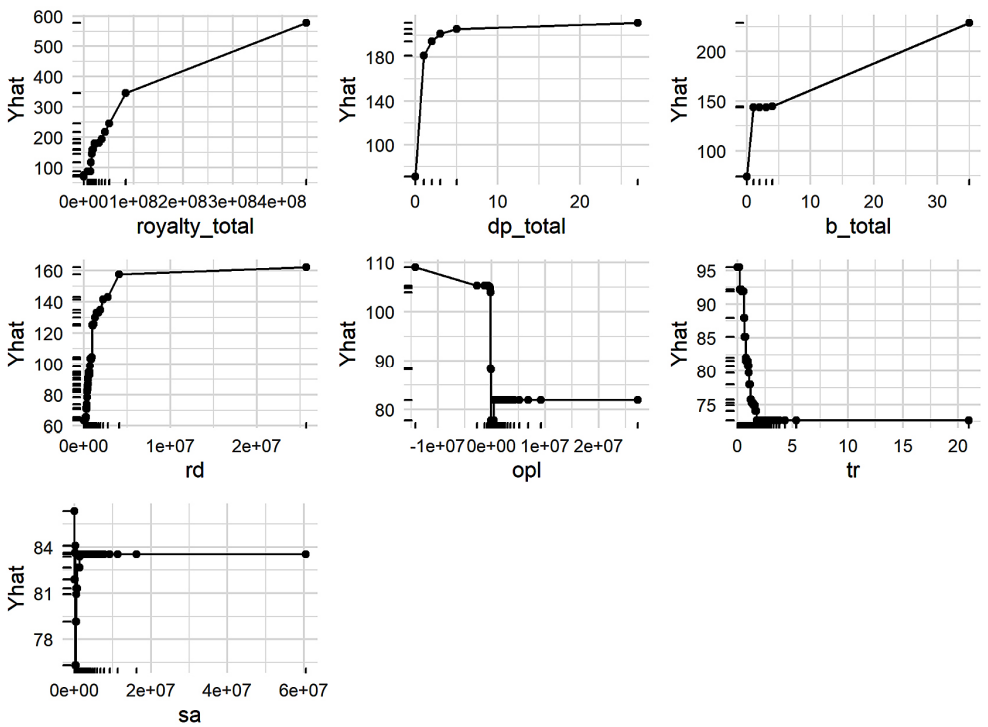
〈Table 5〉 Variables in 'Very High' Cluster, 'High' Cluster, and 'Middle' Cluster

Variable	Definition	Category	Importance	Cluster
royalty_total	Total Royalty receipt	Research Performance	19.766	Very High
dp_total	Total Domestic Patent	Research Performance	10.434	High
b_total	Total number of business application case	Research Performance	5.923	High
rd	Research and Development Expenses (asset+cost)	Business Performance	5.642	High
da_n_1	Domestic patent application dummy(previous year)	Research Performance	1.137	Middle
b_n_1	Business application dummy(previous year)	Research Performance	0.983	Middle
opl	Operation profit	Business Performance	0.861	Middle
sci_total	SCI publication dummy	Research Performance	0.806	Middle
tr	Total asset turnover	Business Performance	0.581	Middle
sa	Selling and Administrative Expense	Business Performance	0.465	Middle

특히 연구성과지표 중에는 연구성과의 사업화 성과를 나타내는 지표들의 순열 중요도가 높았다. 최고(最高)중요도군과 고(高)중요도군 4개 변수 중 3개 변수인 기술료 총합(Royalty_total), 국내특허총계(dp_total), 사업화총계(b_total)는 모두 기술개발의 성과를 사업화하여 얻은 실적을 나타내는 지표이다. 중(中)중요도군에도 연구성과지표는 3개 지표가 포함되었는데, 그 중 2개 지표인 전년도 사업화 유무(b_n_1) 및 전년도 특허 출원 유무(da_n_1)가 연구결과의 사업화 성과를 나타내는 지표였다, 연구개발 프로젝트의 학술적인 성과를 나타내는 SCI 논문기재 유무(SCI publication) 변수 역시 중(中)중요도군에 포함되었지만, 순열 중요도는 0.806에 그쳤다. 반면 기업경영지표 중 연구개발비 1개 지표만이 최고(最高)중요도군 및 고(高)중요도군에 포함되었는데, 이 지표는 통상적인 기업경영지표 보다는 연구개발과 관계가 높은 지표이다. 그 외 3개 기업경영지표는 중(中)중요도군에 포함되었으나, 순열 중요도가 1.0에 미치지 못하였다. <Table 5>의 결과는 중소기업의 연구개발 프로젝트 성과의 사업화 관련 지표가 정부 연구개발 지원금 규모 예측오차 축소효과가 큰 독립변수임을 보여준다.

4.2.3 변수 영향 분석

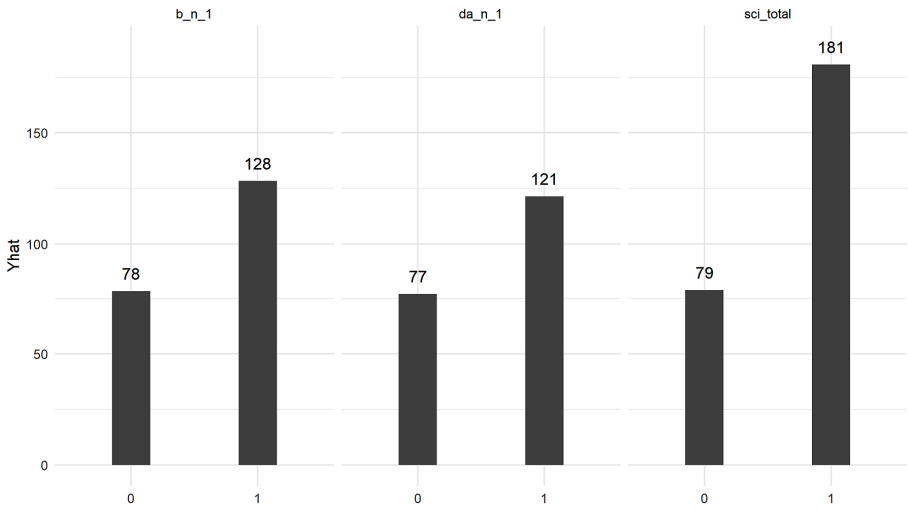
<Table 4> 및 <Table 5>의 순열 중요도 분석결과는 개별 변수가 연구개발 지원금 추정의 정확도에 미치는 영향의 크기에 대한 정보는 제공하지만, 개별 변수가 지원금 규모의 증감에 주는 영향에 대한 정보는 제공하지 못한다. 이를 보완하기 위해서 본 논문에서는 개별 변수의 부분의존도와 SHAP 값을 분석하였다. 부분의존도는 여타 독립변수의 영향 이외에 개별 독립변수가 종속변수 추정치에 미치는 추가적인 영향(Marginal Effect)을 나타내는 지표이고, SHAP 값은 여타 독립변수와의 상호작용을 모두 고려하여 개별 독립변수가 종속변수의 추정치에 미치는 영향을 나타내는 지표이다. 부분의존도는 개별 변수와 여타 독립변수 간 상관관계가 약할 경우에 개별 변수의 영향 파악에 유리하고, SHAP 값은 독립변수 간 상관관계가 있을 경우에 개별 변수의 영향 파악에 유리하다. 부분의존도 분석 및 SHAP 값 분석은 <Table 5>에 수록된 10개 변수들에 대해서만 수행하였다.



〈Figure 3〉 Partial Dependence Plot(PDP): Continuous Independent Variable(unit: million(Yhat), thousand, count, %(independent variable))

부분의존도는 식 (1)에서 추정한 조건부평균함수에서 특정 변수의 값을 고정하고, 나머지 변수의 값에 대해서 평균을 취한 값으로 정의된다[24]. 본 논문에서는 평가표본을 사용하여 부분의존도를 계산하였다. 연속변수인 독립변수는 평가표본 내에서의 100분위 값에 대해서 부분의존도를 구하였고, 이항변수인 독립변수는 개별 독립변수의 값에 대해서 부분의존도를 구하였다. 학습표본을 사용하여 부분의존도를 계산하여도 결과는 유사하였다.

<Figure 3>은 <Table 5>에 수록된 10개 독립변수 중 7개 연속변수의 평가표본의 100분위 값과 각 100분위 값에서 계산한 부분의존도의 관계를 보여준다. <Figure 3>의 그래프의 하단의 눈금은 100분위 값의 분포를 보여준다. 눈금이 조밀한 구간일수록 실측치가 많은 구간이다. <Figure 3>에서 알 수 있듯이 연구개발성과지표가 좋고, 연구개발 투자가 많은 기업들의 정부 연구개발 지원금 추정치가 높은 경향이 존재하였다. 반면 영업이익(opl)과 자기자본회전율(tr)이 높은 기업들은 연구개발 지원금 추정치가 낮은 경향이 존재하였다. 특히 영업이익은 0을 경계로 연구개발 지원금 추정치가 급격하게 하락하는 현상이 발견되었다. 판매비와 관리비(sa)의 부분의존도는 값이 낮은 구간을 제외하고는 거의 변화하지 않았으며, 특정한 경향을 나타내지도 않았다. <Figure 4> 역시 연구개발성과지표가 높은 기업들은 연구개발 지원금 추정치가 높은 경향을 보여준다. 특히 SCI 논문 발표 성과가 있는 경우에는 연구개발 지원금 추정치가 크게 증가하였다.



〈Figure 4〉 Partial Dependence Plot(PDP): Binary Independent Variable(unit: million (Yhat))

<Table 6>은 <Table 5>에 수록된 10개 독립변수의 증감에 따른 부분의존도의 증감을 수록하였다. 7개 연속변수에 대해서는 독립변수의 평가표본 1분위 증가에 따른 부분의존도의 증감을 독립변수의 100분위 값에 대해서 계산하여 평균을 취하였으며, 3개 이산변수에 대해서는 독립변수가 1의 값을 가질 경우의 부분의존도에서 0의 값을 가질 경우의 부분의존도를 차감하였다. <Table 6>은 <Figure 3> 및 <Figure 4>에서 관찰한 독립변수의 값과 부분의존도간의 관계를 정량적으로 요약하여 보여준다. 실제로 기업경영지표인 영업이익(opl) 및 총자산회전율(tr)은 독립변수의 값이 증가함에 따라 부분의존도가 평균적으로 감소하였고, 판매비 및 관리비는 독립변수의 값이 1% 증가하여도 부분의존도의 변화는 -0.028에 그쳐 0에 가까웠다. 반면 연구성과지표 6개 및 연구개발비는 모두 독립변수의 값이 증가함에 따라서 부분의존도가 뚜렷하게 증가하였다.

부분의존도는 관심의 대상이 되는 변수를 고정하고, 다른 독립변수에 대해서는 평균을 취한 값을 사용한다. 따라서 개별 변수가 여타 변수와의 상호작용을 통해서 추정치에 미치는 영향을 파악하기에는 적합하지 않다. 이를 보완하기 위해서 타 독립변수와의 상호작용을 포괄하여 개별 독립변수가 추정치에 미치는 영향을 평가하는 SHAP 값을 사용하였다.

〈Table 6〉 Average Marginal Variation of Partial Dependence Plot

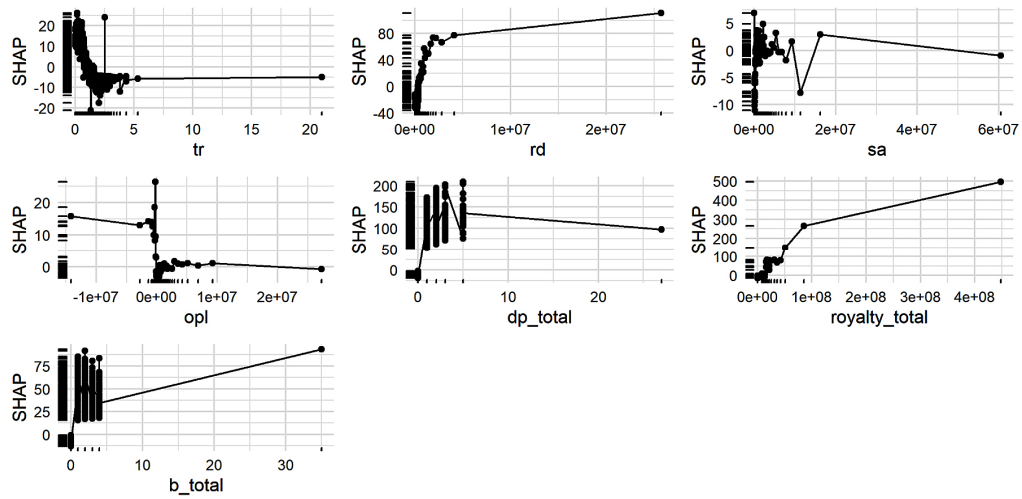
Variable	d(Partial Dependence)/d(X)
royalty_total	2.786
dp_total	1.360
b_total	0.716
rd	0.949
opl	-0.275
tr	-0.230
sa	-0.028
Variable	d(Partial Dependence)
da_n_1	44.369
b_n_1	50.192
sci_total	102.115

SHAP 값은 추정대상인 조건부평균 함수를 독립변수의 선형결합으로 근사하는 특정한 설명모형(Explanation model)에서 각 독립변수에 부여되는 가중치의 값이다. 개별 독립변수에 부여되는 가중치는 해당 독립변수가 포함될 수 있는 독립변수의 부분집합이 종속변수 추정치에 갖는 영향력을 그 독립변수가 포함될 수 있는 모든 부분집합에 대하여 계산하고, 이를 개별 독립변수마다 가중평균한 값을 사용한다. 직관적으로 SHAP 값은 특정 관측치에서 독립변수가 특정한 값을 가졌기 때문에 평균적인 조건부평균 함수의 값(추정치의 평균)과 대비하여 발생하는 특정 관측치에서의 조건부평균 함수의 값(추정치)의 격차를 나타낸다[22, 24]. 본 논문에서는 의사결정나무를 응용한 모형에서 SHAP 값을 빠르게 계산하기 위해서 개발한 TreeSHAP[22]를 사용하였다.

설명모형은 특정 관측치 근방에서만 추정대상과 근사한 값을 갖기 때문에, 선형 설명모형은 개별 관측치마다 각 독립변수에게 다른 가중치를 부여한다. SHAP 값은 선형인 설명모형에서 독립변수에 부여되는 가중치으므로, 모든 관측치에서 모든 독립변수는 상이한 SHAP 값을 갖는다. 개별 독립변수의 SHAP 값의 절대값을 취하여 모든 관측치에 대해서 평균을 계산하면, 이는 여타 독립변수와의 상호작용을 고려한 개별 독립변수가 추정치에 미치는 평균적인 영향을 보여준다. 이 값은 독립변수의 중요성을 평가하는 대안적인 척도로 사용된다. <Table 6>에 수록된 10개 변수가 SHAP 절대값의 평가표본 평균이 가장 높은 10개 변수와 일치하였다. 그리고 평가표본에서 계산한 평균 SHAP 절대값으로 3개 군집으로 나누는 군집분석을 적용하면 평균 SHAP 절대값이 높은 상위 2개 집합에 속한 변수가 9개 선정되었는데, 이 변수들은 모두 <Table 6>에 수록된 변수였다. 4개 이상의 군집으로 구분하는 경우에도 상위 2개 집합에 속한 변수는 8개로 안정적으로 유지되었는데, 이 변수들 역시 <Table 6>에 수록된 변수에 속하였다. 따라서 순열중요도로 평가한 변수의 중요도와 평균 SHAP 절대값으로 평가한 변수의 중요도는 유사하였다. <Table 7>은 <Table 6>에 수록한 변수들의 평가표본에서의 평균 SHAP 절대값과, 평가표본에서의 평균 SHAP 절대값을 기준으로 3개 군집('High', 'Middle', 'Low')로 분할하였을 경우 소속된 군집을 보여준다.

〈Table 7〉 Variable with High Average Absolute SHAP Value

Variable	Importance	Importance Cluster	average absolute SHAP	SHAP Cluster
royalty_total	19.766	Very High	23.383	High
dp_total	10.434	High	21.980	High
b_total	5.923	High	14.577	High
rd	5.642	High	19.327	High
da_n_1	1.137	Middle	3.724	Middle
b_n_1	0.983	Middle	2.284	Middle
opl	0.861	Middle	3.160	Middle
sci_total	0.806	Middle	1.678	Low
tr	0.581	Middle	5.742	Middle
sa	0.465	Middle	2.705	Middle

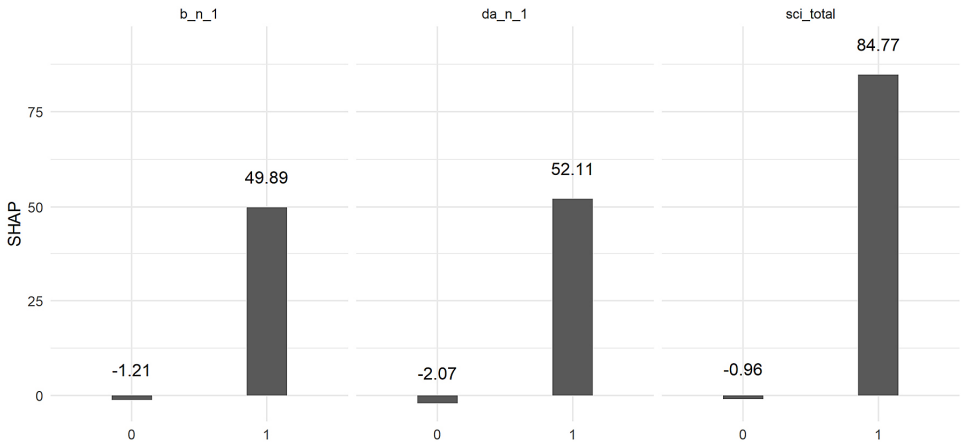


〈Figure 5〉 SHAP Value: Continuous Independent Variable(unit: million(SHAP), thousand, count, %(independent variable))

모든 관측치에서 모든 독립변수의 SHAP 값이 존재하므로, 개별 독립변수의 값의 증감에 따른 SHAP 값의 변화를 추적할 수 있다. <Figure 5>는 <Table 6>에 수록된 10개 독립변수 중 7개 독립변수의 평가표본의 100분위 값을 갖는 관측치에서의 SHAP 값과 독립변수의 100분위 값 사이의 관계를 보여준다. <Figure 3>의 경우와 같이 그래프 하단의 눈금이 조밀할수록 실측치가 많은 구간이다. SHAP 값이 양의 값을 가지면 독립변수가 해당 값을 갖는 관측치에서 그 독립변수의 값이 연구개발 지원금 추정치를 증가시키는 경향이 있음을 의미하고, SHAP 값이 음의 값을 가지면 독립변수가 해당 값을 갖는 관측치에서 그 독립변수의 값이 연구개발 지원금 추정치를 감소시키는 경향이 있음을 의미한다.

<Figure 5>에서 국내특허총합계(dp_total), 기술료 총합계(Royalty_total), 사업화 총합계(b_total) 3개 연구성과지표는 전 구간에서 SHAP 값이 양의 값을 가졌다. 이는 이상 3개 지표는 전 구간에서 연구개발 지원금 추정치를 증가시키는 경향이 있음을 보여준다. 연구개발비(rd)는 값이 낮은 구간에서 SHAP 값이 음의 값을 가졌지만, 값이 증가하면서 SHAP 값이 증가하여 전반적인 구간에서 양의 값이 유지되었다. 이는 연구개발비는 전반적으로 연구개발 지원금 추정치를 증가시키는 경향이 있음을 보여준다. 따라서 <Figure 3>에서 발견한 연구성과지표가 높고 연구개발 투자가 많은 기업이 보다 많은 연구개발 지원금을 받는 경향은 <Figure 5>에서도 확인된다.

총자산회전율(tr)은 연구개발 투자와는 반대로 값이 작을 경우에는 SHAP 값이 양의 값을 가졌지만, 값이 증가할수록 SHAP 값이 급격하게 감소하여 전반적으로 SHAP 값이 음의 값을 가졌다. 그리고 영업이익(opl)은 음의 값을 갖는 구간에서는 SHAP 값이 양의 값을 가졌지만 양의 값을 갖는 구간에서는 SHAP 값이 0과 가까운 값을 가졌다. 이는 총자산회전율은 증가할수록 연구개발 지원금 추정치를 낮추는 경향이 있음을 보여주고, 영업이익은 음의 값을

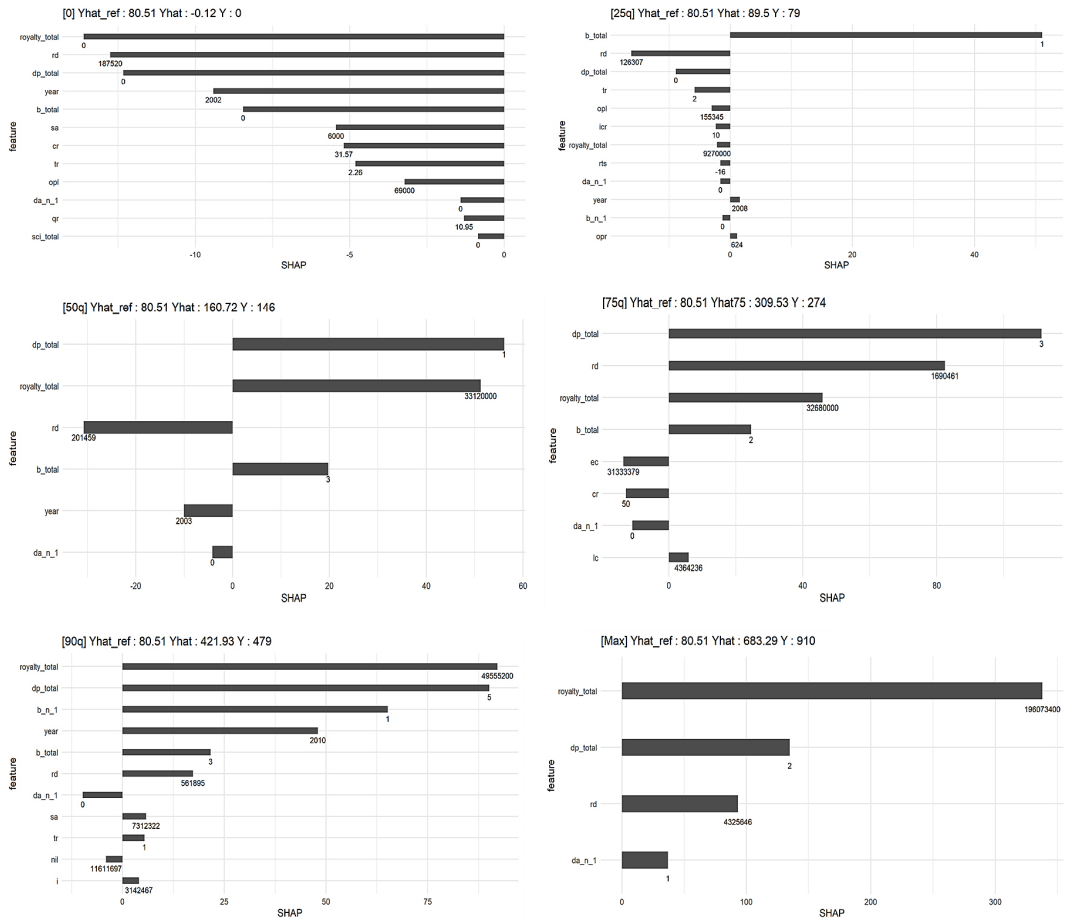


〈Figure 6〉 Average SHAP Value: Binary Independent Variable(unit: million)

가지면 연구개발 지원금 추정치를 높이지만, 양의 값을 가지면 연구개발 지원금 추정치에 영향을 주지 않는 경향이 있음을 보여준다. 따라서 <Figure 3>에서 발견한 자기자산 회전율이 높고 이윤이 높은 기업은 연구개발 지원금 추정치가 낮은 경향은 <Figure 5>에서도 확인된다. 반면 판매비 및 관리비(sa)는 양의 값을 갖는 구간에서 SHAP 값이 전반적으로 양의 값을 가졌다. 이는 판매비 및 관리비 지출은 연구개발 지원금 추정치를 높이는 경향이 있음을 의미한다. 이러한 현상은 SHAP 값에서만 나타난다.

<Figure 6>은 <Table 6>에 수록된 10개 독립변수 중 3개 이항변수가 0의 값을 갖는 평가표본에서 해당 독립변수의 SHAP 값의 평균과 1의 값을 갖는 평가표본에서 해당 독립변수의 SHAP 값의 평균을 보여준다. 0의 값을 갖는 표본의 SHAP 값은 평균적으로 음의 값이었고, 1의 값을 갖는 표본의 SHAP 값은 평균적으로 양의 값이었다. <Figure 6>은 SCI 출판 실적이 있거나(sci_total), 전년도 국내특허출원 실적이 있거나(da_n_1), 전년도 사업화 실적이 있을 경우(b_n_1) 연구개발 지원금 추정치가 증가하는 경향이 있으며, SCI 출판 실적이 없거나, 전년도 국태특허출원 실적이 없거나, 전년도 사업화 실적이 없는 경우에는 연구개발 지원금 추정치가 감소하는 경향이 있음을 보여준다. 이 역시 <Figure 4>와 일관된 결과이다.

SHAP 값은 모든 관측치에서 개별 독립변수의 영향을 평가할 수 있기 때문에, <Figure 5>와 같이 특정 관측치에서의 추정치와 추정치의 평균값 간의 격차에 개별 독립변수가 어느 정도 공헌했는지 파악할 수 있다. 이와 같은 분석을 수행하면 특정한 관측치에서도 <Figure 5>과 <Figure 6>에서 확인된 관계가 나타나는지 파악할 수 있다. 본 논문에서는 종속변수의 값이 0([0]), 평가표본에서 0 이상인 지원금의 1사분위값(25%)([25q]), 중위값(50%)([50q]), 3사분위값(75%)([75q]), 상위 90%([90q]), 최대값([max])을 갖는 관측치 중 각각 추정오차가 가장 작은 값 6개 값을 선정하였고, 개별 관측치에서 SHAP 값과 해당 관측치의 추정치 간의 관계를 파악하였다. <Figure 7>은 그 결과를 보여준다. <Figure 7>에서 막대그래프에 기입된 값은 각 독립변수의 관측치에서의 값이다.



〈Figure 7〉 SHAP Force Plot at Six Sample Observation in Test Set(unit: million)

<Figure 7>에서는 추정치의 값이 높은 관측치에서는 연구성과지표 및 연구개발비의 값이 크고, SHAP 값은 양의 값을 가지며, 추정치의 값이 낮은 관측치에서는 연구성과지표 및 연구개발 투자의 값도 작고 SHAP 값도 작은 경향이 관찰된다. 그리고 종속변수가 0 인 경우([0]) 및 양의 값을 갖는 종속변수의 1사분위 값인 경우([25q])에는 영업이익(opl)이 양의 값을 가져서 SHAP 값은 음의 값을 갖는 현상을 확인할 수 있었다. 또한 양의 값을 갖는 종속변수의 3사분위 값([75q])에서는 판매비 및 관리비(sa)가 양의 값을 갖고 SHAP 값도 양의 값을 갖는 현상을 확인할 수 있었다. 6개 모든 관측치에서 연구개발성과지표의 SHAP 값은 기업경영지표의 SHAP 값보다 절대값이 큰 경향이 있었다. 이러한 경향들은 모두 <Figure 5>와 <Figure 6>에서 확인된 경향과 일관된다.

부분의존도와 SHAP 값을 이용하여 개별 변수가 정부 연구개발지원금 추정치이 미친 영향을 분석한 결과 다음과 같이 정리할 수 있다. 첫째, 연구개발지원금 추정치에 미치는 영향이 큰 변수가 순열중요도도 높았다. 둘째, 중요도가

높은 10개 독립변수 중에서는 연구개발성과지표 및 연구개발비가 연구개발지원금 추정치에 미치는 영향이 기업경영지표가 연구개발지원금에 미치는 영향보다 크게 나타났다. 셋째, 연구개발성과지표 및 연구개발비 지출이 큰 기업은 연구개발지원금 추정치가 커지는 경향이 존재한다. 넷째, 기업경영지표 중 영업이윤이 0 이상의 값을 갖거나 총자산 회전율이 높은 값을 가지면 연구개발 지원금 추정치가 작아지는 경향이 존재한다.

4.3 시사점

기계학습을 이용한 정부 중소기업 연구개발 지원금 추정 결과가 주는 시사점은 다음과 같다. 첫째, 중소기업 연구개발 지원금은 비선형성이 강하여 값이 큰 구간에서는 추정오차가 커진다. 둘째, 연구성과지표 및 연구개발비 지출은 연구개발 지원금의 추정 오차를 축소하는 효과가 크다. 셋째, 연구성과지표가 좋고 연구개발비 지출이 큰 기업의 연구개발 추정치가 큰 경향이 있다. 넷째, 기업경영지표 중에서는 영업이익이 작고 자기자산 대체율이 낮아 기업경영성과가 부진한 기업의 연구개발 추정치가 큰 경향이 있다. 이상의 분석 결과는 현재 중소기업 연구개발 지원금은 연구개발지표가 좋고 연구개발투자가 많은 기업에게 우선적으로 배분되는 반면, 기업경영성과가 좋은 기업에게는 우선 배분되지 않고 있음을 시사한다. 따라서 본 연구의 결과는 현재 중소기업 연구개발 지원금 배분 방식은 기업이 연구개발 지원금 획득을 목적으로 연구성과를 제고할 유인은 제공하고 있지만, 같은 목적으로 경영성과를 제고할 유인을 제공하는 기능은 부족함을 시사한다.

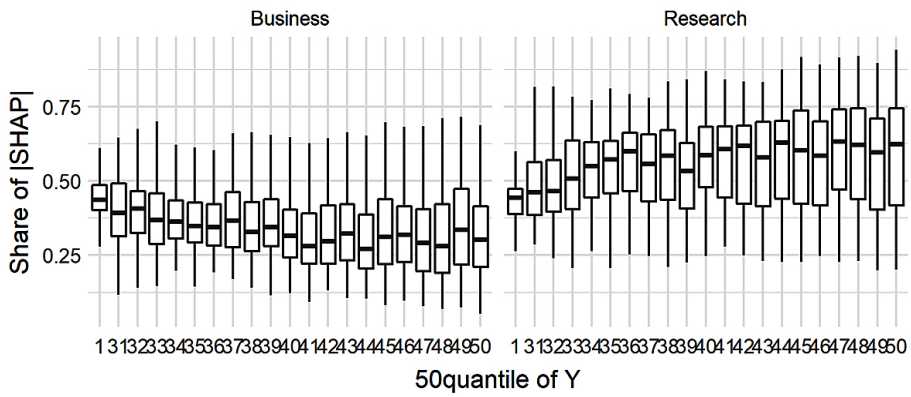
첫째, <Figure 2>로부터 알 수 있듯이 정부의 중소기업 연구개발 지원금은 규모가 커지면 추정의 불확실성이 커지는 경향이 있다. 본 논문에서 기업경영지표 및 연구성과지표를 폭넓게 활용하고서도 지원금 규모가 큰 구간에서는 추정오차를 줄이기 어려웠다. 이러한 현상은 기업경영지표와 연구성과지표의 제곱값, 세제곱값 및 초월함수 값을 독립변수로 사용하여도 관찰되었으며, <Table 2> 및 <Table 3>에 수록한 바와 같이 그래디언트 부스팅 모형보다 더 자유도가 높은 신경망 모형을 사용하였을 때도 관찰되었다. 따라서 현재의 독립변수들의 정보로는 지원금 규모가 큰 구간에서는 정확한 추정이 어렵다고 할 수 있다. 이러한 결과는 다음과 같이 해석될 수 있다. 정부의 중소기업 연구개발 지원금 책정 방식이 지원금이 적을 때는 기업의 연구성과지표 및 연구개발 투자액을 기준으로 이뤄질 수 있으나, 지원금 규모가 커지면 정부는 전형적인 규칙과는 별개의 책정 방식을 채택하고 있다는 점을 시사한다. 이는 결과적으로 정부 연구개발 지원금 신청을 희망하는 기업들에게 일관성을 찾지 못하게 해 혼란을 줄 수 있다는 점에서 문제가 된다.

둘째, 이러한 불확실성이 존재하는 상황에서도 연구성과지표 및 연구개발 투자액은 연구개발 지원금 추정의 불확실성을 줄이는 효과가 있었다. <Table 6>에서 확인되는 바와 같이 그래디언트 부스팅 모형에서 사용한 8개 연구성과지표 중 6개가 추정오차를 가장 많이 축소한 변수군(群)에 포함되었다. 그리고 <Table 7>에서 확인된 바와 같이 <Table 6>에 수록된 6개의 연구성과지표가 평균 SHAP 절대값으로 평가한 추정치에 대한 영향이 가장 큰 변수군(群)에도 포함되었다. 또한 연구개발비는 추정오차 축소

효과가 네 번째로 컸으며, 평균 SHAP 절대값으로 평가한 추정치에 대한 영향은 세 번째로 높았다. 반면, 분석에 포함된 41개 기업경영지표 중 연구개발비(rd), 자기자본대체율(tr), 판매비와 관리비(sa), 영업이익(opl) 4개만이 추정오차 축소 효과가 큰 변수군(群)에 포함되었다. 그리고 평균 SHAP 절대값으로 평가한 연구비 추정치에 미치는 영향 역시 6개 연구성과지표 중 기술료 총합계(royalty_total), 국내특허총합계(dp_total), 사업화총합계(b_total) 3개 지표 및 연구개발비(rd)는 14.5 이상이었고, 3개 기업경영지표는 5.7 이하였다. 이러한 결과는 연구개발 지원금 배분 방식에 연구성과지표와 연구개발비 지출이 기업경영지표보다 더 큰 영향을 미치고 있음을 시사한다.

이와 같은 결과가 모든 변수를 고려해도 유지되는지를 점검하기 위해서 평가표본의 모든 관측치에서 SHAP의 절대값의 합에서 41개 기업경영지표의 SHAP 값의 절대값의 합, 8개 연구성과지표의 SHAP 값의 절대값의 합이 차지하는 비중을 계산하였다. 그 결과 기업경영지표의 SHAP 값 절대값 합이 차지하는 비중의 평가표본에서의 평균은 41.1%였던 반면, 연구성과지표의 SHAP 값 절대값 합이 차지하는 비중의 평가표본에서의 평균은 48.0%에 해당하였다. 이러한 현상이 소수 이상치(outlier)에서만 나타나는지를 점검하기 위하여 평가표본을 연구개발 지원금의 50 분위로 분할하고 각 분위 내에서 기업경영지표 SHAP 값의 절대값의 합이 전체 SHAP값의 절대값의 합에서 차지하는 비중 및 연구개발성과지표의 SHAP 값의 절대값의 합이 전체 SHAP값의 절대값의 합에서 차지하는 비중의 상자그림을 구하여 <Figure 8>에 수록하였다. <Figure 8>의 왼쪽 그래프는 기업경영지표의 SHAP 값 절대값의 합의 비중의 연구개발 지원금 분위별 분포 추이를, 오른쪽 그래프는 연구개발성과지표의 SHAP 값 절대값의 비중의 연구개발 지원금 분위별 분포 추이를 의미한다. 연구개발 지원금의 규모가 작은 분위에서는 기업경영지표의 SHAP 값의 절대값의 합의 비중과 연구성과지표 SHAP 값의 절대값의 합의 비중이 유사하였지만, 연구개발 지원금의 규모가 점차 증가함에 따라 연구성과지표 SHAP 값의 절대값의 합의 비중이 커지는 경향이 존재하였다. 그리고 31분위 이상에서는 연구성과지표 SHAP 값의 절대값의 합의 비중이 더 큰 경향을 확인할 수 있었다. <Figure 8>은 전반적으로 연구성과지표가 연구개발 지원금 추정치에 미치는 영향은 기업경영지표가 연구개발 지원금 추정치에 미치는 영향보다 컸으며, 이러한 현상은 연구개발 지원금이 증가할수록 심화되었음을 보여준다. 이러한 결과는 본 논문의 분석에서 활용된 모든 기업경영지표와 연구성과지표를 고려한다고 해도, 연구개발 지원금 배분 방식에 연구성과지표와 연구개발비 지출이 기업경영지표보다 더 큰 영향을 미치고 있음을 시사한다.

셋째, 연구개발성과지표가 좋고 연구개발비 지출이 큰 중소기업의 연구개발 지원금 추정치가 상대적으로 큰 경향이 있었다. 연구개발성과 지표 및 연구개발비는 관측치가 증가함에 따라 부분의존도가 증가하였고, 전 구간에서 SHAP 값이 양의 값을 가졌다. 이러한 결과는 현재 정부의 중소기업 연구개발 지원금 배분 방식이 중소기업에게는 연구개발성과지표를 개선하고 연구개발비 투자를 확대하는 유인을 제공하고 있음을 시사한다.



〈Figure 8〉 Absolute SHAP Share: Business vs Research Performance

넷째, 기업경영지표중 영업이익이 음의 값을 갖고 총자산회전율이 낮고 판매비 및 관리비 지출이 큰 기업의 연구개발 지원금 추정치가 상대적으로 큰 경향이 있었다. 총자산회전율의 SHAP 값은 전반적으로 음의 값을 가지고 있었으며, 총자산회전율이 증가하면 SHAP 값이 더 하락하였다. 그리고 총자산회전율이 증가하면 부분의존도 역시 하락하였다. 이는 총자산증가율이 증가할수록 연구개발 지원금 추정치가 감소함을 의미한다. 영업이익은 0 보다 작을 경우에는 부분의존도와 SHAP 값이 양의 값을 갖고 있었지만 0을 기점으로 부분의존도가 급격하게 하락하였고, 0보다 높은 값에서는 부분의존도가 일정하게 유지되었다. 영업이익의 SHAP 값 역시 영업이익이 0보다 작은 구간에서는 양의 값을 가졌지만, 0을 기점으로 급격하게 하락하여 0에 접근하였다. 이는 영업이익이 0 보다 낮으면 연구개발 지원금 추정치가 높지만, 0을 기점으로 급격하게 감소함을 의미한다.

마지막으로 판매비 및 관리비가 작은 구간에서는 SHAP 값이 음의 값을 유지하였지만, 증가할수록 SHAP 값이 증가하여 큰 값에서는 SHAP 값이 양의 값을 가졌다. 이는 판매비 및 관리비가 증가할수록 연구개발 지원금 추정치가 증가함을 의미한다.

총자산회전율은 매출액을 자산총계로 나눈 값이다. 따라서 본 논문에서 수행한 3가지 기업경영지표의 영향 분석 결과는 매출이 적어서 총자산회전율이 낮고, 영업이익은 적고, 판매비 및 관리비 지출은 큰 기업은 연구개발 지원금 추정치가 큰 경향이 있음을 의미한다. 이러한 결과는 현재 정부의 중소기업 연구개발 지원금 배분 방식이 기업의 매출 제고, 이윤 신장, 비용 절감을 제고할 유인이 부족함을 시사한다.

이를 종합하면 현재 정부의 중소기업 연구개발 지원금 배분 방식은 중소기업이 연구개발 지원금 수령 액수 제고를 목적으로 연구성과지표 개선 및 연구개발비 지출 확대를 시도할 유인을 제공하고 있지만, 기업경영실적을 제고할 유인은 부족함을 시사한다.

5. 결 론

본 논문은 정부 중소기업 연구개발 지원금 배분을 결정하는 요인을 기계학습을 적용하여

파악하고자 하였다. 이를 위하여 본 논문은 한국과학기술평가원에서 구축한 중소기업 연구개발 지원금 자료에 그래디언트 부스팅 모형을 적용하여 중소기업 연구개발 지원금 추정모형을 개발하였고, 순열중요도 분석을 시행하여 추정오차 축소 효과가 큰 변수를 선정하였으며, 선정된 변수의 부분의존도 및 SHAP 값을 분석하여 지원금 추정치를 증가시키는 독립변수를 파악하였다. 추정 결과 선형회귀분석 응용모형에 비해서 평균제곱근오차를 7.20% 개선하는 지원금 추정모형을 개발하였고, 연구개발성과지표 및 연구개발비 지출이 추정오차를 줄이는 효과가 큰 독립변수임을 확인하였으며, 연구성과지표가 좋고 연구개발비 지출이 많은 기업의 지원금 추정치가 커지는 경향을 확인하였다. 그리고 총자본회전율이 낮고, 영업이익이 음의 값을 갖고, 판매비 및 관리비 지출이 큰 기업의 지원금 추정치가 커지는 경향을 확인하였다. 본 논문의 결과는 정부의 중소기업 연구개발 지원금 배분 방식이 연구성과 개선 및 연구비 투자 증대 유인은 제공할 수 있지만, 기업경영을 개선할 유인은 부족함을 시사한다.

본 논문의 성과는 네 가지로 정리할 수 있다. 첫째 본 논문은 정부 연구개발 지원금 배분 방식이 지원금 경쟁에 참여하는 기업들에게 제공하는 유인의 파악에 집중했다는 점에서 지원받은 기업의 성과에 분석에 집중한 기존 연구와 차별성이 있다. 둘째, 본 논문은 공시적 가정이 성립되어 있지 않은 정부의 의사결정 과정에 대한 실증분석에 기계학습을 사용하여 시사점을 도출하는 선례를 제공하였다. 셋째, 본 논문은 그래디언트 부스팅 모형을 사용하여 선형회귀분석 모형보다 추정의 평균제곱근오차를 7.20% 축소하는 중소기업 대상 연구개발 지원금 추정모형을 개발하였다. 넷째, 본 논문은 부분의존도 및 SHAP 값을 적극적으로 이용하여 개별 변수의 영향력의 크기 및 방향성을 파악하였다.

학술적으로 본 논문은 기계학습 연구와 정책연구의 결합이라는 점에서 선행연구와 차별성을 갖는다. 본 논문은 정책연구의 관점에서 방법론의 제약으로 연구가 상대적으로 부족하였던 중소기업 연구개발 지원금 배분 방식 연구에 기계학습 방법론을 적용하여 방법론의 제약을 완화하였다. 또한, 본 연구는 기계학습 연구의 관점에서도 최근의 연구성과인 부분의존도와 SHAP 값을 사용하여 기계학습 연구 결과의 정책적인 활용도를 제고하였다. 본 연구에서 개발한 중소기업 연구개발 지원금 추정모형은 중소기업 지원을 담당하는 정책부서에서는 중소기업 지원금 배분 현황을 파악하는 도구로 활용할 수 있으며, 중소기업의 연구개발 지원금 지원 시에 지원금을 예측하고 연구개발 지원금 수령액수를 높이기 위해 필요한 요인을 파악하는 도구로 활용할 수 있을 것으로 기대한다.

반면, 본 논문에서 개발한 모형은 기업경영지표와 연구성과지표들을 폭넓게 활용했고 추정모형 후보군도 다양하게 활용했음에도 불구하고 지원금 추정모형 내에서 지원금이 큰 구간에서 추정오차가 커지는 한계가 있었다. 그리고 본 논문에서 사용한 자료는 2010년 이전의 자료이므로, 최근 10년간의 변화를 반영하지 못하였다는 한계는 존재한다. 단, 본 논문에서 사용한 자료는 국가 연구개발 사업 관련 자료이고, 국가 연구개발 사업의 효율성에 대한 문제제기는 2010년 이래 최근까지 지속되고 있다[18, 26]. 따라서 본 논문의 결과는 최근의 국가 연구개발 사업 관련 연구에도 시사점을 지닌다.

본 논문의 결과는 다음과 같은 두 가지 추가적 연구과제를 제공한다. 첫째, 우선은 지원금이 큰 구간에서의 추정오차를 낮추기 위한 후속연구가 필요하다. <Figure 2>에서 확인되듯이 추정오차는 지원금 규모가 특정한 임계점을 넘어서면 급격하게 확대되는 경향이 있다. 이 경우에는 종속변수를 지원금이 0 인 경우, 지원금이 0 이상이지만 임계점을 넘지 않는 경우, 그리고 지원금이 임계점을 넘는 경우를 나타내는 이산변수로 전환하여 분석하면 보다 추정 성과가 좋은 모형을 개발할 가능성이 있다. 둘째, 본 논문의 결과는 중소기업 연구개발 지원금이 기업성과가 좋은 기업보다는 연구성과가 좋은 기업에게 우선 제공되고 있을 가능성을 시사한다. 정부 연구개발 지원의 궁극적인 목적이 기업의 경영성과 제고임을 감안하면, 연구개발 지원을 통해 경영성과 개선 유인을 제고할 수 있는 제도개선이 필요하다. 구체적인 제도개선 관련 연구는 추후의 과제로 돌린다.

References

[1] Bergstra, J. S., Bardenet, R., Bengio, Y., and Kégl, B., Algorithms for hyper-parameter optimization, NIPS'11: Proceedings of the 24th International Conference on Neural Information Processing Systems, pp. 2546-2554, 2011.

[2] Bloom, N., Reenen, J. V., and Williams, H., "A Toolkit of Policies to Promote Innovation," Journal of Economic Perspectives, Vol. 33, No 3, pp. 163-84, 2019.

[3] Chang, W. H., "Is Korea's Public Funding for SMEs Achieving Its Intended Goals?," KDI Focus, No. 63, 2016. 2. 3.

[4] Choi, J. M., "A Study of the Effects of Government R&D Support on Product Innovation in Small and Medium-sized Enterprises(SMEs): Focusing on the Moderating Effect of Firm Characteristics," Korean Journal of Public Administration, Vol. 56, No. 2, pp. 213-248, 2018.

[5] Cin, B., Kim, Y., and Vonortas, N. S., "The Impact of Government R&D Subsidy on Firm Performance: Evidence from Korean SMEs," Small Business Economics, Vol. 48, No. 2, pp. 345-360, 2017.

[6] Fisher, A., Rudin, C., and Dominici, F., "All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously," Journal of Machine Learning Research, Vol. 20, No. 177, pp. 1-81, 2019.

[7] Friedman, J. H., "Greedy function approximation: a gradient boosting machine," Annals of statistics, Vol. 29, No. 5, pp. 1189-1232, 2001.

[8] Gerath, J., Witten, D., Hastie, T., and Tibshirani, R., An Introduction to Statistical Learning, New York: Springer, 2013.

[9] Hall, B. H. and Lerner, J., Chapter 14-The financing of R&D and innovation, In Handbook of the Economics of Innovation, Vol. 1, pp. 609-639, 2010.

[10] Hong, J. P. and Kim, J. H., "Impacts of Financial Policies for SMEs on Firms Performance: Role of Supplier Network between Large Firms and SMEs," Journal of Korean Economic Analysis, Vol. 21, No. 3, pp. 185-240, 2015.

[11] Ivezić, Ž., Connolly, A. J., VanderPlas, J. T., and Gray, A., Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data. Princeton University Press, 2019.

[12] Ji, M. W., "Did Legal Criteria for Receiving Governmental Support Cause a Negative Effect in Employment Growth of SMEs?: Evidence from the Korean Manufacturing Industry," The Journal of Korean Public Policy, Vol. 17, No. 3, pp. 3-31, 2015.

[13] Jun, B. W. and Choi, E., "Review on Tax Expenditures for Small-and-Mid Sized Firms," Asia Pacific Journal of Small Business, Vol. 37, No. 3, pp. 1-24, 2015.

[14] Kang et al., "An empirical Study on the Impact of Government R&D Investment on SMEs in Korea," Korea Institute of S&T Evaluation and Planning, Report no. 2016-027, 2016.

[15] Kang et al., "Big Data Analysis: Application to Environmental Research and Service II," Korea Environment Institute, 2018.

[16] Kang et al., "Big Data Analysis: Application to Environmental Research and Service," Korea Environment Institute, 2017.

[17] Kim, K. H. and Yang, J. Y., "Government R&D Support and Apply Strategy for SMEs," Regional Industry Review, Vol. 41, No. 3, pp. 299-324, 2018.

[18] Kim, K. W., Kim, J., Shin, J. K., and Hong, S. B., How to Improve the efficiency of Government R&D Investment, Korea Development Institute, 2011.

[19] Ko, H. S., Chung, Y. H., Seo, H. K., and Song, L. K., "A Study on the Effectiveness of the SMEs Consulting Support Project: Focused on Hidden Champion Business Supporting in Daejeon," Asia Pacific Journal of Small Business, Vol. 38, No. 1, pp. 169-188, 2016.

[20] Kuhn, M. and Johnson, K., Applied predictive modeling(Vol. 26), New York: Springer, 2013.

[21] Lee, D. H. and Kim, K. H., "Deep Learning Based Prediction Method of Long-term Photovoltaic Power Generation Using Meteorological and Seasonal Information," The Journal of Society for e-Business Studies, Vol. 24, No. 1, pp. 1-16, 2019.

[22] Lerner, J., Boulevard of broken dreams: why public efforts to boost entrepreneurship and venture capital have failed and what to do about it. Princeton University Press, 2009.

[23] Li, T., Jing, B., Ying, N., and Yu, X., "Adaptive Scaling," arXiv preprint arXiv: 1709.00566, 2017.

[24] Lundberg, S. M. and Lee, S. I., "A unified approach to interpreting model predictions," In Advances in neural information

processing systems (pp. 4765-4774), 2017.

[25] Lundberg, S. M., Erion, G. G., and Lee, S. I., "Consistent individualized feature attribution for tree ensembles," arXiv preprint arXiv:1802.03888, 2018.

[26] Molnar, Christoph. Interpretable Machine Learning, Lulu.com, 2020.

[27] National Assembly Budget Office, Analysis on Government R&D Program : Overview, Seoul, 2019.

[28] OECD, The SME Financing Gap (Vol. I): Theory and Evidence, OECD Publishing, Paris, 2006.

[29] Pyo, H. H. and Choi, H. H., "The Effects of Export Promotion on Korean Manufacturing SMEs' Performance," Kukje Kyungje Yongu, Vol. 24, No. 3, pp. 29-56, 2018.

[30] Strobl, C., Boulesteix, A., Zeileis, A., and Hothorn, T., "Bias in random forest variable importance measures: Illustrations, sources and a solution," BMC Bioinformatics, Vol. 25, No. 8, pp. 1-21, 2007.

[31] Zhao, Q. and Hastie, T., "Causal interpretations of black-box models," Journal of Business & Economic Statistics, DOI: 10.10870/07350015, 2019.

[32] Zúñiga-Vincente, J. A., Alonso-Borrego, C., Forcadell, F. J., and Galán, J. I., "Assessing the effect of public subsidies on firm R&D investment: a survey," Journal of Economic Surveys, Vol. 28, No. 1, pp. 36-67, 2014.

〈부록 1〉 기계학습에 사용된 변수

〈Appendix Table 1〉 List of Variables used in Machine Learning

Category	Variable	Definition	unit	Category	Variable	Definition	unit
Dependent Variable	gov_fund	Government R&D Subsidy	million	Business	car	Capital Adiquacy Ratio	ratio
Business	kk	Revenue	thousand	Business	opr	Operation Profit Growth Ratio	ratio
Business	cr	Current Ratio	ratio	Business	pr	Net Profit Growth Rate	ratio
Business	qr	Quick Ratio	ratio	Business	ri	Return on Investment	ratio
Business	dr	Debt Ratio	ratio	Business	eg	Employment Growth Rate	ratio
Business	fr	Fixed Ratio	ratio	Business	rd	R&D expense (asset +cost)	thousand
Business	icr	Interest Coverage Ratio	ratio	Business	cs	cost of goods sold	thousand
Business	sgr	Sales Growth Rate	ratio	Business	sa	selling and administra-tive expenses	thousand
Business	rts	Total Asset Growth Rate	ratio	Business	as	Total asset	thousand
Business	roe	Return on Equity	ratio	Business	tfaa	fixed tangible asset	thousand
Business	rn	Net Profit Margin	ratio	Business	opl	Operation Profit	thousand
Business	ois	Operating Profit Ratio	ratio	Business	npl	Net Profit	thousand
Business	itr	Inventory Turnover Ratio	ratio	Business	nil	Net Income	thousand
Business	rt	Receivable Turnover	ratio	Business	ec	Equity Capital	thousand
Business	tfa	Tangible Fixed Asset Growth Rate	ratio	Business	oc	Outside Capital	thousand
Business	dc	depreciation	ratio	Business	i	Inventory	thousand
Business	t	taxes and dues	thousand	Business	ie	Interest Expenses	thousand
Research	dp_total	Total domestic patent	number	Business	fe	Financial Expenses	thousand
Research	royalty _total	Total royalty receipt	thousand	Business	dar	Total Borrowing and Bonds payable to Total Asset	ratio
Research	b_total	Total number of business application	number	Business	pc	personnel cost	thousand
Research	op_total	Total patent abroad dumpy	bianry	Business	lc	labor cost	thousand
Research	sci_total	sci publication dummy	binary	Business	fff	Other employee benefits	thousand
Research	dr_n_1	Domestic Patent Registration dummy (previous year)	binary	Business	ii	interest income	thousand
Research	b_n_1	Business application dummy(Previous year)	binary	Business	rc	rent cost	thousand
Research	da_n_1	Domstic Patent Application dummy (previous year)	binary	Business	tr	Total asset turnover	number
				control	year	year of gov_fund disbursement	number

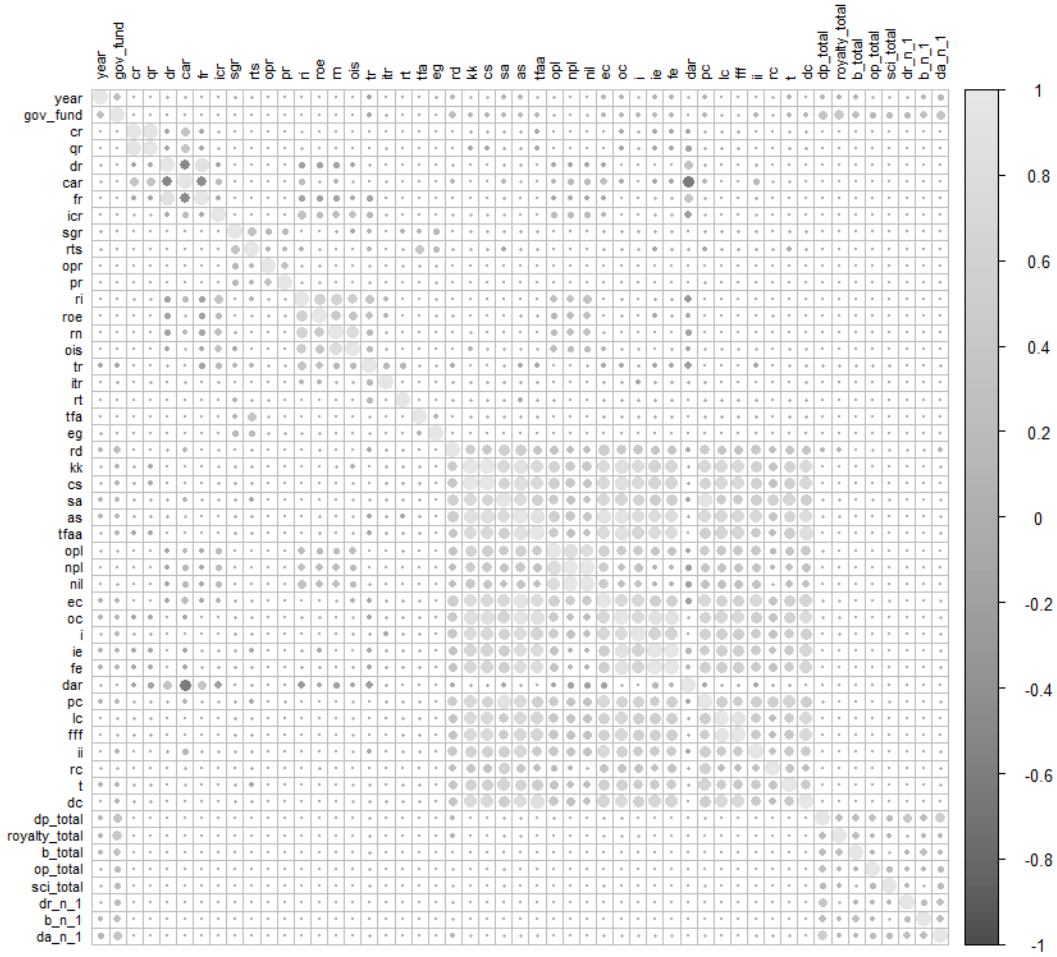
〈부록 2〉 표본통계량

〈Appendix Table 2〉 Sample Statistics of Continuous Variable

Variable	Mean	SD	MAX	MIN	Variable	Mean	SD	MAX	MIN
year	2006	2	2010	2002	cs	9717011	22554701	415000000	38
gov_fund	81	155	911	0	sa	1708184	3385303	76206169	654
cr	260	464	8629	-469	as	11036845	23386202	350000000	6964
qr	212	408	7900	-123	tfaa	3837064	9167557	157000000	-2554
dr	251	460	10073	-8	opl	624075	2024318	27847361	-24826939
car	42	21	108	0	npl	511220	2193306	23202836	-22170278
fr	171	278	6494	0	nil	417930	1916594	19926460	-20839122
icr	7	36	328	-320	ec	4995240	12272719	178000000	-22205084
sgr	60	190	2921	-100	oc	6041225	13024032	217000000	-6713
rts	40	98	2200	-100	i	1464185	3700860	72179719	49
opr	119	402	6355	-100	ie	206066	488599	9510090	1
pr	140	459	6793	-100	fe	221754	534846	10196451	1
ri	5	17	283	-431	dar	35	24	795	0
roe	14	45	1114	-1160	pc	660027	1191419	27047539	21
rn	0	28	300	-317	lc	841044	1841938	34297040	100
ois	-1	31	99	-385	fff	109172	250688	4669501	5
tr	1	1	29	0	ii	56632	161317	2534251	-172
itr	50	126	1334	0	rc	32565	78750	1997048	1
rt	11	22	350	0	t	30935	59409	1495668	1
tfa	68	235	2512	-108	dc	413528	1073522	22556285	7
eg	13	49	1200	-99	dp_total	0	1	46	0
rd	437398	838340	25734436	1	royalty_total	5507892	21662315	812771000	0
kk	11966050	26079417	459000000	2400	b_total	0	1	81	0

〈Appendix Table 3〉 Sample Frequency of Binary Variable

Variable	0	1
op_total	40353	335
sci_total	40298	390
dr_n_1	40159	529
b_n_1	39748	940
da_n_1	39354	1334



〈Apendix Figure 1〉 Sample Correlation Plot

저 자 소 개

강성원	(E-mail: swkang@kei.re.kr)
1996년	서울대학교 경제학과 (학사)
1999년	서울대학교 경제학과 (석사)
2006년	Rudgers University Economics (박사)
2010년	삼성경제연구소 수석연구원
2011년 6월	한국경제연구원 연구위원
2011년 8월～현재	한국환경정책평가연구원 연구위원
관심분야	일반균형모형, 온실가스감축, Machine Learning

강희찬	(E-mail: henrykang@inu.ac.kr)
1998년	고려대학교 농업경제학과 (학사)
2001년	고려대학교 경제학과 (석사)
2006년	오하이오 주립대학교 자원환경경제학과 (박사)
2006년	삼성경제연구소 수석연구원
2012년	한국환경정책평가연구원 연구위원
2014년～현재	인천대학교 경제학과 교수
관심분야	비시장가치평가법, 기후변화 정책