「지방정부연구」 제24권 제4호(2021 겨울): 443-476

The Korean Journal of Local Government Studies, Vol.24 No.4 (2021 Winter)
http://dx.doi.org/10.20484/klog.24.4.18

한국 지방소멸 요인과 극복 방안에 관한 연구: 머신러닝 방법을 통한 탐색*

유 한 별**
탁 근 주***
문 정 승****

국문요약

본 연구는 한국의 시군구 지방자치단체(이하 지자체)의 소멸 위험을 탐색하고, 소멸 위험에 영향을 미치거나 지역의 매력도에 영향을 미치는 요인을 도출한다. 이 분석을 위해, 국가통계포털에서 변수 자료를 수집·구성하고, 각 지자체의 인구이동을 반영한 개선 소멸위험지수 도출한다. 해당 소멸위험지수로부터 각 지자체의 소멸위험등급을 생성한 후 해당 등급을 벡터 변환한 목표값(target value)과 지자체 별 독립변수를 활용하는 머신러닝 분석 모델을 구축한다.

머신러닝 모델은 GBM, RF, XGB 등을 활용하며, 보팅(voting), 앙상블(ensemble) 등 모델의 성능을 향상시키는 방법으로 지자체의 소멸위험등급을 예측하고 분류한다. 이 결과를 통해 소멸위험이 높은 지자체를 도출하며, 해당 지자체의 소멸 위험에 영향을 주는 요인을 판별한다.

분석결과, 본 연구에서 구축한 머신러닝 모델의 예측(prediction) 성능은 약 90% 내외였으며, 68개의 지자체가 소멸 위험이 가장 높은 등급으로 측정되었다. 이러한 소멸위험에 영향을 주는 요인은 인구사회학적 요인, 경제·산업적 요인, 문화·의료 시설 등의 편의 요인 등이 영향을 주는 것으로 확인되었다.

결론적으로 본 연구를 통해 향후 소멸위험에 처한 지자체가 지방소멸을 극복하기 위해서는 이러한 경제·산업적 요인에 먼저 집중이 필요하며, 해당 요인이 극복된다면 문화·의료시설을 확충하여 지역 매력도를 높이는 것이 중요하다고 볼 수 있다.

주제어: 지방소멸, 지방소멸요인, 지방소멸위험, 지역매력도, 머신러닝

* 본 연구는 2020년 빅데이터 청년인재 양성사업(데이터 청년 캠퍼스) 프로젝트 평가에서 최우수상을 수상한 논문을 수정 작성하였습니다. 논문의 수정과 발전에 도움을 주신 모든 심사위원님들께 깊은 감사의 말씀을 올립니다.

본 논문은 2017년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임.(NRF-2017S1A3A2067636).

** 제 1저자
*** 공동저자
**** 공동저자

Ⅰ. 서론

2020년 1월 국내 코로나 바이러스(COVID-19, 이하 코로나)의 확산이 시작되면서 '사회적 거리두기'가 시행되었으며, 전국적으로 코로나 확산의 여파가 지속되면서 경제적인 타격으로 이어졌다. 자영업자 및 중소기업 등이 코로나 확산으로 영업에 타격을 입어 종업원, 근로자를 줄이고 무급휴가를 장려하는 등의 조치를 취하게 되면서 자영업 및 중소기업의 생태계도 위협을 받고 있다. 특히, 이미 2020년 3월부터 전년 동월 대비 취업자 수가 감소하였고, 2020년 7월 기준으로 자영업자는 554만 명으로 지난해보다 13만 명이 감소하는 추세를 보였으며, 제조 중소기업의 공장가동률이 1월부터 6월까지 지속적으로 70% 이하를 기록하면서 금융위기 이후 가장 큰 타격을 입고 있다.[1)]

이러한 취업률 저하, 자영업과 중소기업의 어려움은 지방 소도시의 자영업·중소기업 종사자들의 소득 저하로 이어졌다. 이 상황을 극복하기 위해 고용상황이 지방보다 덜 나쁜 수도권으로 청년·중장년층이 이동하면서 지방소멸의 가속화를 불러일으킬 가능성을 높이고 있다. 실제로 한국고용정보원의 보고서에 따르면 2020년 3월~4월 수도권 유입 인구 중 20대의 비중이 75.5%였으며, 전국 228개 시군구 기준 소멸위험지역이 19년 5월 93개에서 20년 4월 기준 105개로 12곳 증가하였다고 파악되었다(한국고용정보원, 2020). 이러한 추세는 코로나 감염병의 확산으로 가속화되었으나, 이미 일본의 지방소멸의 모습과도 비슷하며, 행정·제도적인 영향을 일본과 미국으로부터 받아온 우리나라는 과거 일본과 인구 구조도 비슷한 양상을 보이고 있기 때문에 지방소멸에 대한 논의는 앞으로 지속될 것으로 보인다(하연섭, 2020).

일본은 이미 1990년 저성장과 경제 불황, 2000년 초 인구감소 등을 경험하며 지방소멸에 대한 위기의식을 가지고 이를 대응하기 시작하였다(임보영 외, 2018). 하지만 경기 회복은 더디게 진행되었으며, 2011년 일본 국토교통성의 「국토의 장기전망」보고서와 2014년 「마스다보고서」에 따르면 지방 도시 인구감소가 심화되어 일본 전체의 인구감소로 이어질 수 있고, 인구감소가 심화되면 2050년에는 일본 전 국토의 3분의 2이상이 인구가 절반이상 감소하고, 국토의 5분의 1은 사람이 살지 않는 곳이 될 수 있다는 분석 결과를 내었다(이호상, 2016). 이러한 상황에서 일본은 지자체 주도적 규제 완화, 특구계획 등으로 직접적으로 국토 공간 지역 활성화를 정책을 시도하였다. 또한 '도시재생긴급정비지역'으로 특정지역을 지정하여 대도시와 지방중핵도시의 규제 완화를 진행하여 도시재생사업을 진행하였다(임보영 외, 2018).

하지만 규제 완화와 특구지정은 지방소도시보다는 거점 중핵도시 위주였기 때문에 중핵도시가 아닌 지방소도시는 쇠퇴문제에서 벗어나기 어려웠고, 지방 도시들의 불만이 일어났기 때문에 이러한 불만을 수용하여 '마을·일·사람 창생법'을 제정하고 '국토그랜드디자인 2050(2014)', '국토형

1) 고용동향 브리프, 한국고용정보원, 2020. vol.1
코로나發 '줄폐업'에…무너지는 자영업 생태계, 이데일리, 2020.08.20.
(https://www.edaily.co.kr/news/read?newsId=01308726625869288&mediaCodeNo=257),
제조 중소기업 공장가동률 5개월째 70% 미달…금융위기 후 처음, 매일경제, 2020.08.13.
(https://www.mk.co.kr/news/economy/view/2020/08/831144/)

성계획(2015)' 등을 정비하여 공간을 재편하고자 하였다(김은혜·박배균, 2016; 이호상, 2016; 임보영 외, 2018). 이를 통해 대도시·중소도시의 공간을 축소하여 교통·정보·서비스 등의 활성화와 규모의 경제를 실현하고자 하였으며, 청년층 수입 확보와 경력 단절 지원, 노동 환경 개선, 해외 인재 수용 등의 인구 정책을 펼치는 등의 노력을 진행하고 있다.

상기한 바와 같이 일본은 다각적인 정책적 노력을 진행하였으나, 여전히 지방소멸의 위험에 직면하고 있다. 마찬가지로 앞서 진행된 연구에 근거하여 우리나라 또한 이러한 단계를 밟아갈 가능성이 크고, 앞서 서술한 바와 같이 코로나의 영향으로 지방소멸이 가속화 되고 있다. 이러한 상황을 인식한 국회는 지방소멸 문제를 해결하고자 '인구소멸위기지역 지원 특별법 제정안' 등 법안 3건을 추진하고, 공청회를 진행하는 등 지방소멸에 대한 위기를 인식하고 대처하려는 노력을 진행하고 있다.[2)]

이러한 시점에서 본 연구에서는 우리나라의 지방소멸 위기를 인식하고 현 상태를 진단하기 위해, 지방자치단체 별 지방소멸 위험지역을 분류함과 동시에 지방소멸에 미치는 영향 요인들을 구체화하고자 한다. 이를 통해 지방소멸에 미치는 특정한 영향 요인들이 각 지방소멸위험지역에 어떻게 영향을 미치고 있는지를 구체적으로 파악한 후 지방소멸의 극복가능성에 대한 논의를 진행한다. 이러한 논의를 바탕으로 향후 지방소멸 가능성을 낮추기 위한 지역의 취약 요인 극복 방안에 대한 결론과 정책적 함의를 제공하고자 한다.

Ⅱ. 이론적 논의

1. 한국의 지방소멸 위기와 일본의 지방소멸

지방소멸은 일본에서 먼저 대두되었고, 최근 우리나라에서도 관심이 높아지고 있다. 일본과 우리나라에서 나타나고 있는 지방소멸이 무엇인지에 대해 살펴보기 위해서는 지방소멸의 정의와 현황에 대한 논의가 필요하다. 먼저 지방 소멸에 대한 정의를 살펴보고 논의를 심화하기로 한다. 2014년 5월 발표된 일본창성회의의 보고서인 「성장을 이어가는 21세기를 위하여: 저출산 극복을 위한 지방활성화 전략(ストップ少子化·地方元気戦略, 이하 마스다 보고서)」에서는 인구 유출의 지속으로 인구가 사라질 가능성이 높은 지역을 소멸 지역으로 보고 이를 지방소멸로 칭하였다(박승현, 2017). 결국 지방소멸은 말 그대로 '지방이 소멸한다(없어진다).'는 사전적인 뜻이 될 수 있고, 그 의미는 지방의 인구가 대도시, 중핵도시 등으로 유출되며 해당 지역의 인구 공동화(人口 空洞化) 현상이 나타는 것으로 볼 수 있다.

마스다 보고서에서는 일본 지방의 인구유출이 지속된다면 일본의 지방 지역 중 896개가 사라질 가능성이 높다고 예측하였다. 일본은 해당 보고서 출판 이후 지속적으로 정책적인 개입을 통해 공

2) '지방소멸 위기' 특별법 국회 공청회 개최, KBS NEWS, 2020.0818(http://news.kbs.co.kr/news/view.do?ncd=4519288)

간을 축소하고 공간 네트워크를 강화하여 지역연결망 구축과 도쿄 중심의 극점사회를 극복하려고 하였으나 여전히 도쿄 중심 사회를 극복하지 못하였고 향후 낮은 출생률이 지속된다면 도쿄중심의 극점사회가 더욱 심화될 수 있다는 전망이 나오고 있다(박승현, 2017). 이러한 이유로 지방소멸의 논의는 일본에서도 상당히 지속적으로 논의되고 있으며, 그 관심도 또한 높다. 우리나라의 지방소멸은 일본과 많이 닮아있다. 일본이 도쿄 중심이라면 우리나라는 서울에 정치, 경제, 사회, 문화가 집중되어있다.

서울은 이미 조선의 한양으로부터 입지를 공고히 했고, 광복과 6·25 전쟁 이후 청와대, 국회 등 정부기관이 상주하는 도시로 관습적으로 정치·행정의 중추적 역할을 담당하는 대한민국의 수도 역할을 해왔다. 이는 '신행정수도의건설을위한특별조치법위헌확인, 헌재 2004. 10. 21. 2004헌마554 등, 공보 제98호, 1095' 소(訴)의 판결 전문을 통해서도 알 수 있으며, 향후에도 서울이 점하는 타지역 대비 우위적 지위는 지속될 것으로 보인다. 결국 최근 서울 중심의 수도권의 인구는 이미 비수도권의 인구를 뛰어 넘었다.[3] 코로나의 여파이긴 하나 지방소멸에 대한 위기의식을 고취시키고, 지방자치단체와 국회의원들이 자신들의 지방의 소멸 가능성에 대한 논의를 하게 하는데 충분하였다.

결국 근래에 '지방중소도시의 지방소멸 방지를 위한 특별법 제정'에 관한 법제화 움직임이 국회에서 시작되었으며, 지방 공청회를 진행하는 등 지방 소멸 논의가 활발히 진행되고 있다. 이러한 법제화를 통해 지방중소도시들에 대한 소멸 위험을 절감하고 자생력을 키우려는 시도를 하고 있는 것이다. 지방자치단체의 자생력을 키우기 위한 정부와 국회의 움직임이 전개되고 있는 시점에서 현재 우리나라의 지방소멸에 대한 예측은 마스다 보고서에서 제시한 지방소멸지수를 통해 제시되고 있으며, 해당 지수는 재생산가능인구와 노인인구의 비율을 통해 소멸 가능성을 제시하고 있다.[4]

하지만 기존 논의되고 있는 지방소멸위험 지수는 인구감소를 출생과 사망의 차이에서 나타나는 자연 증감에만 주목하고 있다는 한계가 지적되어오고 있다(정성호, 2019). 또한 한 시점의 지역소멸위험도 측정은 인구 증감 원인을 고령인구의 증가와 젊은 여성인구의 감소로 한정짓는다는 지적도 존재한다. 즉, 일자리 창출이나 취학 등의 요인으로 인한 사회적인 인구이동 변수를 고려하지 않고 있다는 점에서, 오늘날 국내 지방의 인구감소 문제가 대도시로의 인구이동에서 비롯된다는 면을 간과하고 볼 수 있다(이상호, 2016). 본 연구에서는 지방소멸위험 지수에 대한 개념에 대하여 지역의 인구 이동에 대한 정보를 포함시키기 위해 한 지역의 전입과 전출이라는 요소를 추가한 후, 새로운 지방소멸위험 지수를 산출하고, 등급을 세분화하고자 한다.

본 연구에서는 지방소멸 위기를 인식하고, 지방소멸위험지역에 대한 예측과 그 극복에 대한 논의를 진행하고자 한다. 자세히 서술하면, 소멸위험지수에 근거하여 도출된 지방소멸위험지역들을

3) 수도권인구, 비수도권 사상 첫 추월…20대 직장·학교 찾아 서울로, 연합뉴스, 2020.06.29.(https://www.yna.co.kr/view/AKR20200629063600002)

4) 소멸위험지수 = 한 지역의 20~39세 여성인구 수/해당 지역의 65세 이상 고령인구 수, 마스다 보고서에서는 해당 값이 0.5 미만이면 소멸 위험지역으로 정의함.

위험수준별로 분류하여 해당 지역들의 소멸위험지수에 미치는 영향요인을 도출하기 위한 분석을 진행하여 결과를 제시한다. 이 분석결과를 통해 지방소멸에 영향을 주는 요인들을 도출하고, 해당 요인들로부터 지방소멸을 극복할 수 있는 정책적 함의를 제시하기로 한다.

2. 지방 매력도(amenity)와 지방정부기능을 통한 소멸 요인 탐색

지방소멸은 대도시, 중핵도시 집중화로 인해 지방의 인구가 유출됨으로써 발생할 수 있음을 앞서 서술하였다. 대도시, 중핵도시 집중화는 경제·정치·행정적 집중화를 야기하고, 이러한 집중화는 해당 분야들에서 대도시, 중핵도시와 지방중소도시의 격차를 형성하게 만든다. 이러한 격차는 상대적으로 지방의 매력도를 감소시키게 되며, 매력도 감소가 결국 인구 유출로 이어진다. 지방의 매력도 감소가 결국 지방소멸로 이어질 수 있음을 이러한 메커니즘(mechanism)을 통해 알 수 있다(McNulty et al, 1984; Logan & Molotch, 1987; Gottlieb, 1994; Choi, 2012).

지방 매력도는 경제적인 요인뿐만 아니라, 소위 어메니티(amenity)라는 요소로 구성될 수 있으며, 지방 주민들의 주거이전(인구 유입과 유출)에는 근린어메니티, 환경어메니티, 도시어메니티가 영향을 미칠 수 있다(강병수, 2014). 근린어메니티는 의료서비스, 주택가격, 교육의 질(초·중·고등학교의 질), 범죄 예방 등이며, 도시어메니티는 백화점·대형마트, 대학의 질, 도서관, 극장 등으로 볼 수 있다. 더하여 환경어메니티는 대기의 질, 상수도 수질, 공원 및 오픈 스페이스 등 다양한 요인들로 볼 수 있다.

이러한 요인들은 지방의 매력도를 높이는 특성으로 작용하며 특정 지역 외에서 해당 지방으로 들어오는 유입 인구, 유출 인구 즉, 인구이동에 영향을 미칠 수 있다(최유진, 2017). 특히, 지역의 매력도가 증가하게 되면 관광자원으로 활용하거나, 어메니티 중심의 경제 활성화로 일자리창출, 부동산 가치 상승, 지방자치단체의 세수확장, 인구 유입으로 인한 수요 증가와 산업 활성화 등의 다양한 효과를 창출할 수 있다. 본 연구에서는 지역 어메니티를 구성하는 요소들에 주목하여 이러한 요소들이 지역 소멸 위험 가능성에 미치는 영향을 규명하고자 머신러닝 분석을 통해 해당 요소들의 영향력을 확인하고자 한다.

이를 위해 지역 어메니티에 영향을 미칠 수 있는 요소들에 대한 논의를 진행하여야 한다. 앞서 서술한대로 지역 어메니티는 근린, 환경, 도시 어메니티 등이 있는데 이러한 어메니티는 지방정부의 기능이 원활이 이루어지면 확보된다고 할 수 있다. 그 이유는 근린 시설의 경우 의료, 주택, 교육 등 사회 기반 시설 요소를 다루고 있기 때문에 해당 분야에 대한 정부의 세출, 투자, 정책 등 다양한 요소가 영향을 미칠 수 있으며, 도시와 환경 어메니티 또한 마찬가지로 해당 분야에 대한 정부의 역할이 중요하다. 이러한 면에서 본 연구에서는 어메니티를 이끌어 낼 수 있는 각 지방정부의 기능적 요소를 자료(data)로 구성하여 소멸 위험지수에 이러한 지방정부의 기능적 요소가 미치는 영향에 대한 논의를 진행한다.

지역의 어메니티에 영향을 미칠 수 있는 지방정부의 기능적 요소는 정부 세출예산체계를 나누어 OECD, IMF 등에 보고하는 SNA기준 정부기능분류(이하 COFOG; Classification of the Functions

of Government)의 기준을 따라 논하기로 한다(김성자 외, 2016). 다시 말해, 지방정부의 기능적 요인을 COFOG 분류에 따라 구성하여 해당 요인들이 지방소멸위험에 미치는 영향을 분석하고자 한다. 본 연구에서는 이러한 분석을 위해 COFOG 세출 10개 분야[5]에 따라 지방정부에 대한 각 분야별 자료(data)를 목표값(target value)에 영향을 미칠 수 있는 특성(feature)으로 구성하고, 소멸위험을 목표값으로 하여 지방의 매력도에 영향을 줄 수 있는 지방정부의 기능적 요인이 소멸위험에 미치는 영향을 분석하기로 한다.

3. 선행연구 분석

지방소멸은 현재 우리나라가 직면한 문제로 인식되어 다양한 시각에서 연구되고 있다. 이러한 연구들은 지방과 소멸에 대한 고찰을 진행한 논문의 형태와 소멸 관련 사실과 극복 방안을 연구하는 연구 보고서 등으로 나눌 수 있다. 이러한 논문과 보고서의 주제 또한 다양한데, 먼저, 이미 지방의 소멸을 인식하고 선제적으로 대응에 나섰던 일본의 사례들에 대한 선행연구들을 살펴보기로 한다.

박승현(2017)은 마스다 보고서에 대한 상세한 소개와 함께 일본 아베내각의 지방창생과 마스다 보고서의 지방소멸에 대한 논의를 진행한다. 해당 연구에서 도쿄 일극에 집중하는 문제가 마스다 보고서를 통해 제시되었음을 서술하며, 중앙이 지속적으로 팽창하고 지방 살리기가 지방의 문제로 여겨지고 생활안정을 목표로 하지 않고 인구이동에 초점을 맞춘 출산육아지원, 고용안정 등의 정책을 펼친다면 지방소멸을 극복할 수 없음을 제시하고 있다.

이정환(2017) 또한 마스다보고서를 중심으로 논의한 연구를 통해 마스다보고서가 지방소멸을 제시하고 있고, 해당 보고서가 지방의 커뮤니티, 지방 거주민들의 의사를 반영하지 못하고 있으며 인구 재배치를 중심으로 논하고 있다는 것에 대한 한계를 지적하며 향후 공동체에 대한 고민이 있어야 함을 제시한다. 이외에도 정성호(2019) 또한 마스다보고서의 논의를 통해 우리나라의 저출산·고령화 문제를 함께 다루었으며, 지방소멸위험지수에 대한 문제점을 지적하며 지방소멸위험지수만으로 지방의 소멸위기를 판단하는 것은 부적절함을 제시하였다.

김순은(2017), 이기배(2017), 하동현(2017), 임보영 외(2018), 이건웅(2020) 등은 인구소멸에 대응하는 일본의 정책을 연구하여 일본의 국토공간전략으로 압축·연계 전략(국토구조 축소 및 네트워크)을 검토하거나, 희망출산율 실현, 인구유출 방지전략, 일자리 연계, 지역 사업 창출, 대학 개혁, 정부관계 기관 지방이전 등 다양한 정책적 수단에 대한 연구를 진행하였다. 이러한 연구들은 일본의 사례를 통해 우리나라에 필요한 정책적 수단에 대한 연구를 진행함으로써 향후 우리나라의 인구소멸을 극복하기 위한 정책적인 함의를 제시하였다.

하혜수(2017)는 이러한 연구와 더불어 지방소멸을 맞이한 시점에서 지방자치를 재검토하는 연

5) COFOG 분류에 따른 분야별 기능별 분류구조로 10개 분야는 1. 일반 공공행정, 2. 국방, 3. 공공질서 및 안전, 4. 경제활동, 5. 환경보호, 6. 주거 및 지역사회 시설, 7.보건, 8. 휴양, 문화, 종교, 9. 교육, 10. 사회보호가 있다.

구를 진행하여 다양성 이론에 따라 지방정부 별로 차등적인 다양성과 혁신아이디어 창출이 지방자치발전에 기여할 수 있으며, 문화이론에 따라 지역문화의 수준에 따라 지방분권 수준을 달리 해야 함을 주장하며, 대도시 지역은 이전 재원 축소 및 지방세 확대를 통한 지방분권을 강화하고, 과소지역에 대해서는 지역문화, 자치역량을 고려하여 낮은 수준의 자치를 시행해야 한다고 보았다.

더하여 김동완(2015), 정성호, 홍창수(2018), 정성호(2019), 강동우(2019) 등의 연구에서는 국내 사례에 대한 검토를 통해 소멸 위험 지방자치단체가 실제로 소멸위험을 직면하고 있음을 실증적으로 제시하거나, 소멸 위험을 줄이기 위한 균형발전, 고용 안정성 증대 등을 제시하면서 국내 지역발전 정책에 대한 정책적인 함의를 제시함으로써 지방소멸을 극복하고자하는 논의를 제시하였다.

앞서 서술한 연구들에서 살펴볼 수 있듯이, 다양한 연구들이 지방소멸에 대해 연구하고 있으나, 실증적으로 지방소멸을 분석하기 보다는 마스다 보고서에서 제시한 바를 비판적으로 검토하고 수용하려는 움직임을 보이고 있으며, 일본의 정책적 사례를 우리나라에 적용하여 지방소멸을 극복해야 한다는 연구들과 우리나라의 현 상황을 분석한 연구들이 다수였다. 하지만 이러한 연구들에도 불구하고 지방정부의 소멸위험에 영향을 미칠 수 있는 요인들을 변수로 구성하여 어떠한 요인이 실제로 지방정부 소멸에 영향을 미치는지에 대한 실증적인 연구는 찾기 어려운 실정이다.

본 연구에서는 이러한 측면에 주목하여, 지방정부의 기능을 반영할 수 있는 변수들을 설정한 후, 해당 변수들을 지방자치단체 별로 본 연구에서 진행하고자 하는 시간적 범위 내에서 시계열적으로 수집하여 실제로 어떠한 요인이 지방자치단체의 소멸위험에 영향을 줄 수 있는지 실증적으로 분석해보기로 한다. 이러한 연구를 위해 다음 절에서 연구설계를 논의하고 연구를 진행해 보기로 한다.

Ⅲ. 연구설계

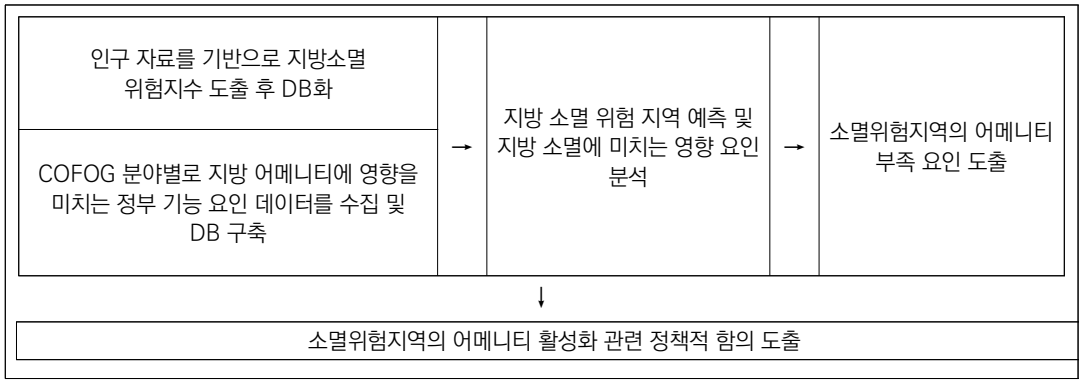
본 연구에서는 앞서 이론적 논의에서 제시한 바와 같이 지방소멸에 대한 논의를 진행하기 위해 지방소멸에 미치는 영향요인, 소멸위험등급예측에 대한 분석을 차례대로 진행한다. 이를 위해 먼저 공간적으로 전국의 지방자치단체는 시군구 단위를 기준으로 하고, 시간적으로 2015년~2018년의 자료를 수집한다. 다음으로 인구 자료를 기반으로 현재 재생산 인구와 노령 인구만으로 도출되는 지방소멸지수를 현행보다 구체화 한 후 해당 지방소멸지수를 4분위수로 나누어 각 시군구에 대한 소멸위험등급을 도출한다. 해당 소멸위험등급을 도출한 후 MariaDB 프로그램을 이용하여 소멸위험등급 및 이에 영향을 미칠 수 있는 지방정부의 기능적 요인들에 대한 데이터베이스를 구축한다. 이후 데이터베이스에 구축된 자료들을 활용하여 Python 3.7, Jupyter Notebook 프로그램을 통해 분석을 진행한다.

분석과정은 다음과 같다. 먼저 탐색적 데이터 분석(이하 EDA: Exploratory Data Analysis)을 실시하여 결측치, 이상치 등을 확인한다. EDA를 마친 후, 데이터 전처리(data preprocessing), 특성 공학(feature engineering)을 진행하기 위해 상기한 동일 프로그램을 활용하고, 마찬가지로 numpy,

pandas, sklearn, matplotlib 등의 패키지를 활용하여 분류 모델을 구축·분석한다. 머신러닝 방법을 적용하기 위해 2018년 이전의 자료(data)를 훈련자료(train set)로, 2018년 자료를 시험자료(test set)로 적용하여 해당 분류 모델의 성능 측정 등을 진행한다. 여러 분류 모델 중 앙상블(ensemble) 기법으로 가장 적합한 분류 모델을 선택하고, 선택된 분류 모델을 기준으로 이후 특성들의 중요도를 도출하고 특정 지역에 미치는 특정 요인들을 도출하여 지역의 소멸가능성을 높이는 특성에 대한 논의를 진행한다.

모델의 성능 강화를 통해 훈련자료로 시험자료의 지방소멸을 정확히 예측·분류하는 방법을 통해 해당 소멸예측에 어떠한 요인이 영향을 미치는지 파악할 수 있으며, 이러한 과정을 통해 향후 지방소멸위험지역에 미치는 영향요인을 근거로 본 연구에서는 정책적 함의를 최종적으로 제시하기로 한다. 이러한 연구형태를 도식화하여 〈그림 1〉 과 같이 연구 진행 모형을 제시할 수 있으며, 해당 연구 모형을 좀 더 상세히 설명하기 위하여 정형 데이터를 활용한 정량적 분석인 분류 모형에 대한 설명 및 설계에 대한 논의를 제시하기로 한다.

〈그림 1〉 연구과정 도식화



1. 소멸위험지수 재설계 및 데이터 수집

본 연구의 정량 분석 설계를 위해 먼저 머신러닝 모델의 지도학습(supervised learning)을 위해 목표값(target value)로 사용할 소멸위험지수를 구성한다. 이를 위해 2015년부터 2018년까지의 각 시군구 지역의 인구를 수집하여, 소멸위험지수를 도출한다. 이 과정에서 본 연구에서는 마스다 보고서에서 제시한 각 지역의 노인 인구로 가임 여성 인구를 나누는 일반적 소멸위험지수의 한계를 인식하고, 소멸위험지수에 대한 개선을 진행하여 분석하고자 한다.

마스다 보고서에서도 소멸위험지수에 대한 논의를 가임 여성의 감소를 제시하면서 가임 여성의 감소가 지방소멸 가능성을 높인다고 제시했을 뿐 소멸의 도달 시기와 해당 소멸위험지수의 정확성 등에 대한 논의는 거의 없다(정성호, 2019). 따라서 본 연구에서는 기존 소멸위험지수의 한계에 대하여 인식하고, 기존 소멸위험지수를 개선하고, 지역의 인구 이동을 반영하기 위해 소멸위험

지수를 개선하여 다음 〈식 1〉과 같이 소멸위험지수를 개선하였다.

〈식 1〉 개선 소멸위험지수 계산식

개선 소멸위험지수 = *(소멸위험지수^2) + log(전입인구/전출인구)*

상기한 〈식 1〉의 소멸위험지수 제곱값을 이용하여 기존 소멸위험지수에 가중(weight)을 주었으며, 전입전출비에 자연로그를 설정한 이유는 해당 값의 분포를 정규분포에 가깝게 구성하기 위한 조치이다.[6] 이 식을 통해 개선된 소멸위험지수를 도출하고, 해당 소멸위험지수를 4분위 수로 나누어 등급화하여 분류 모델의 목표값(target)으로 활용한다.

더하여, 이론적 논의 절에서 지역 어메니티에 영향을 미칠 수 있는 정부기능별분류에 대한 논의를 진행하였다. 정부기능별분류는 앞서 서술한 바와 같이 10개 분야로 구성될 수 있는데 구체적으로 10개 분야는 일반공공행정, 국방, 공공질서 및 안전, 경제활동, 환경보호, 주거 및 지역사회 시설, 보건, 휴양·문화·종교, 교육, 사회보호로 구분된다. 해당 분류를 기반으로 하여 목표값에 영향을 줄 수 있는 특성(feature)을 구성한다. 이러한 특성과 목표값의 구성을 요약하면 아래의 〈표 1〉과 같다.

〈표 1〉 특성에 따른 변수의 구성

특성	자료구성(feature)	특성	자료구성(feature)
목표값 (target value)	'개선 소멸위험지수[7]로 도출된 분류(A~D 등급)'	주거 및 지역사회건설 (9)	'하수도보급률', '상수도보급률' '교통문화지수', '운전행태영역', '교통안전영역', '보행행태영역', '1인당 자동차등록대수', '도시지역면적','주택 수'
일반 공공행정 (15)[8]	'혼인건수', '조혼인율', '합계출산율', '평균연령', '남녀성비', '인구증가율', '당해 년 총 인구', '전년총인구', '천 명당 외국인 수', '주민등록인구', '출생아수', '가구 수', '일반공공행정예산비중', '일반공공행정분야예산액', '토지거래면적'	보건 (12)	'병원 수', '보건소 수', '상급종합병원 수', '약국 수', '요양병원 수', '의원 수', '종합병원 수', '치과병원 수', '한방병원 수', '천 명 당 의료기관병상수', '총 병상 수', '보건분야예산액'
국방 (0)	국가 단위에서 지출이 있는 특성으로 본 연구는 지방 정부를 대상으로 하여 해당 지표를 특성변수에서 제외함.	휴양·문화·종교 (2)	'십만 명 당 문화기반시설 수', '문화기반시설 수'
		교육	'교원 1인 당 학생 수',

6) 자연로그(log)를 적용하지 않은 경우, 값이 한쪽으로 몰리는 문제를 확인하여 차후에 극단적인 값들의 영향이 소멸지수에 상당히 미칠 수 있음을 예상하였으며, 이를 통해 소멸위험등급의 분류가 제대로 진행되지 않을 수 있는 여지를 막기 위해 해당 조치를 취하였다.

		(14)	'재적 학생 수', '교원 수', '유치원교원 수', '유치원 수', '유치원아 수', '천 명 당 사설학원 수', '사설학원 수', '초등 교원 수', '초등학생 수', '유치원학급 당 학생 수', '초등학교학급 당 학생 수', '중학교학급 당 학생 수', '고등학교학급 당 학생 수'
공공질서 및 안전 (0)	'교통사고', '화재', '범죄', '자연재해', '생활안전', '자살', '감염병'의 각 등급[9]		
경제활동 (26)	각 사업[10]의 사업체 수, 종사자 수	사회보호 (11)	'고위험음주율', '비만율', 'EQ.5D(건강상태 표준화)', '주관적건강수준인지율', '인구 십만 명 당 자살률', '건강보험 적용인구 현황', '노인 천 명 당 노인여가복지시설 수', '노인여가 복지시설 수', '60세 이상 주민등록인구', '사회복지예산비중', '사회복지분야예산액'
환경보호 (2)	'녹지지역면적', '녹지지역면적비율'		
자료 범위	시간적 범위: 2015년 ~ 2018년 공간적 범위: 전국 228개 지방자치단체를 대상으로 함		
변수 출처	국가통계포털(KOSIS, http://kosis.kr/)		

위 〈표 1〉에서 제시한 바와 같이 국가통계포털(KOSIS)을 통해 다양한 특성변수들을 구성하여 목표값에 대한 영향을 머신러닝 방법을 이용하여 분석하기로 한다. 목표값은 각 지방자치단체의 소멸위험지수를 〈식 1〉에 근거하여 연속형 변수 상태로 도출한 후 최소, 최대값을 기준으로 4개의 범주로 등급화(4분위)하여 범주형 변수로 전환하는 방법으로 구성한다. 범주형 변수로 전환된 각 지역의 소멸위험등급(A~D등급)을 목표값으로 하여 〈표 1〉에서 제시한 지방정부의 10개 기능적 분야 변수들의 영향력을 확인하기 위하여 국방, 공공질서 및 안전을 제외한 8개 분야에 따라 특성값(feature) 자료를 구성하고, 해당 8개 분류 내 모든 자료 특성을 독립변수(특성값)로, 소멸위험등급을 종속변수(목표값)로 하여 분류 모델 머신러닝을 진행하기로 한다.

2. 데이터 탐색과 핸들링(EDA, data preprocessing, feature engineering)

상기한 바와 같이 자료를 구성한 후 연구에서 사용할 자료에 대한 정제를 진행하기로 한다. 자

7) 〈식 1〉을 통해 인구이동을 반영한 소멸위험지수이다.

8) 해당 특성에 해당하는 특성변수들의 개수를 뜻한다.

9) 2018년도를 기준으로 데이터를 수집, 구성하였으나, 머신러닝 분석에서 타 지표와의 균형성의 문제와 2017년도 이전 KOSIS 데이터의 부재의 한계로 해당 지표를 최종모델의 특성변수에서 제외한다.

10) 농업, 광업, 제조업, 건설업, 도소매업, 운수창고업, 숙박음식점업, 전기가스증기및공기조절공급업, 정보통신업, 금융보험업, 수도하수폐기물원료재생업, 부동산업, 전문과학기술서비스업 등을 반영한다.

료의 정제는 먼저 자료 전처리(data preprocessing)과 특성 공학(feature engineering) 등을 진행하여, 결측치 확인, 타겟변수 설정, 다중공선성 검사, 지표비교, 파생변수 추가 등을 진행한다. 자세한 내용은 추후 서술하기로 한다.

1) 결측치 확인

자료의 결측치 확인은 통계 자료의 구성이 완벽하지 않음을 전제로 각 지방자치단체의 특정 특성의 값이 비어있는 경우(missing data)를 탐색하기 위해 진행한다. 본 연구에서는 전국 228개 지방자치단체에 대한 분석을 진행하며, 결측치가 있는 경우, 해당 지방자치단체 행에 대한 제외(drop)는 진행하지 않고, 해당 지방자치단체의 주변 지방자치단체, 상급 지방자치단체를 고려하여 적정한 값으로 대체한다. 자세히 서술하면, 결측치에 대한 확인을 진행함으로써 특정 변수(index)의 값의 특정 행(row) 값이 비어있는 경우를 탐색하고, 해당 지방자치단체의 중분류 수준에서 같은 행정 구역의 평균으로 대체하거나 지역의 특성을 고려하여 변경할 수 있는 값으로 대체한다.

2) 목표값(target value) 도출

본 연구에서는 앞서 서술한 바와 같이 먼저 〈식 1〉을 통해 소멸위험지수를 도출한 후 해당 소멸위험지수를 활용한 시각화를 진행한다. 지역별 소멸위험지수 시각화를 통해 각 지역이 얼마나 소멸위험에 처해있는지에 대해 확인할 수 있다. 이후 소멸위험지수를 4등급(A~D 등급)으로 나누어 범주형 변수로 구성한다. A~D 등급은 최소값과 최대값을 기준으로 4등급 급간에 따라 특정 값에서 나뉘는 지점(natural break)을 지정하여 등급을 나누며, 이 소멸위험등급을 분류 모델의 목표값으로 설정한다.

3) 다중공선성 확인

본 연구에서는 〈표 1〉에서와 같이 10개 분야 중 8개 분야의 하위 특성변수를 구성하여 분류모델에 반영한다. 목표값을 제외하고, 8개 분야의 독립변수(특성값)로 활용하고자 한 변수는 총 91개로 해당 변수들을 2015년부터 2018년까지 수집하여 분석에 활용하기로 한다. 해당 변수들은 각 분야의 하위 특성으로 구성·수집된 변수들이기 때문에 각 특성변수들 간에 상관관계가 높아서 발생하는 다중공선성이 발생할 수 있다. 다중공선성(multicollinearity)은 특성변수들 간에 선형 상관관계(linear relation)가 있는 경우에 발생할 수 있다. 다중공선성 문제는 회귀분석 등의 모델 파라미터(model parameters)의 신뢰성에 문제로 이어질 수 있다(Aylin, 2010). 이러한 이유로 특성변수 간 다중공선성을 먼저 VIF(Variance Inflation Factor) 분석 결과로 확인한다. VIF 검정은 다중회귀모형에서 독립변수 간 상관관계를 측정하는 척도로 수식은 아래 〈식 2〉와 같이 검정할 수 있다(Aylin, 2010).

〈식 2〉 VIF 검정 일반식

$$VIF_i = \frac{1}{1-R_i^2}, \text{ for } i = 1,2,...,k$$

본래 회귀모형, 로짓모형 등에서는 변수 간 상관관계가 높은 변수들(10 이상의 값) 경우, 다중공선성을 고려하여 해당 변수를 적절히 제외한다. 하지만 본 연구에서는 머신러닝 분류모델을 활용하여 변수 간 상관관계에 영향을 줄이고, 최대한 많은 독립변수들을 고려하여 분류모델의 정확도를 높이고자 하였다. 이는 기존 연구들에서 머신러닝 분류모델이 다항로짓모델[11]보다 독립변수 간 복잡성과 종속변수에 대한 효과를 좀 더 정확히 측정할 수 있음과 동시에 다중공선성의 영향이 적음을 확인하였기 때문에 본 연구에서도 로짓모델과 머신러닝 모델의 정확도를 먼저 확인하여 지방소멸을 정확히 예측할 수 있는 모델을 활용하기로 한다(Xian-Yu, 2011; Tang et al, 2015; 이영호, 홍성연, 2019). 즉, 머신러닝을 활용한 분류모델은 다중공선성이 높은 변수들도 반영하여 설명력을 높일 수 있는 가능성이 있다(김인호, 이경섭, 2020). 따라서 본 연구에서도 다중공선성에 대한 고려를 하지만 앞서 제시한 변수들을 최대한 반영하여 모델의 성능을 비교하면서 변수를 반영하는 방식으로 활용하기로 한다.

4) 분야별 지표 비교

앞서 결측치, 목표값, 다중공선성 등에 대한 탐색과 분석을 진행한 후 특성변수들 간의 특성 중요도(feature importance)에 대한 탐색을 정확하게 진행하기 위해 각 분야별로 변수에 대한 탐색을 진행한다. 이 과정에서 변수별 시각화를 통해 각 분야의 특성변수에 대한 이해와 이상치(outlier)에 대한 탐색, 지역별 특성 분포, 지역 간 특정 특성 차이, 연도별 특성 변화 대한 탐색 등을 진행한다.

3. 분류 모델 제시와 연구 모델 도식화

본 연구에서는 앞서 자료 전처리를 통한 자료 분석과 정제가 완료된 후에 해당 특성변수(특성값)와 소멸위험등급(목표값)을 활용한 분류 모델을 구성한다. 분류모델에는 여러 가지 방법이 있으나, 본 연구에서 활용한 방법에 대해서 간단히 서술하면 아래와 같다. 모델 설명 전에 모델의 과적합을 방지하기 위해 본 연구에서는 데이터 셋(data set)을 분할하여 기본 모델을 구축하고, 하이퍼파라미터(hyper parameter) 선택을 위해 RandomizedSearchCV 모듈을 이용하여 최적화한 후 K-fold 교차검증을 통해 모델의 검증 성능의 신뢰도를 높이고자 하였다. 이러한 작업 후 아래에서

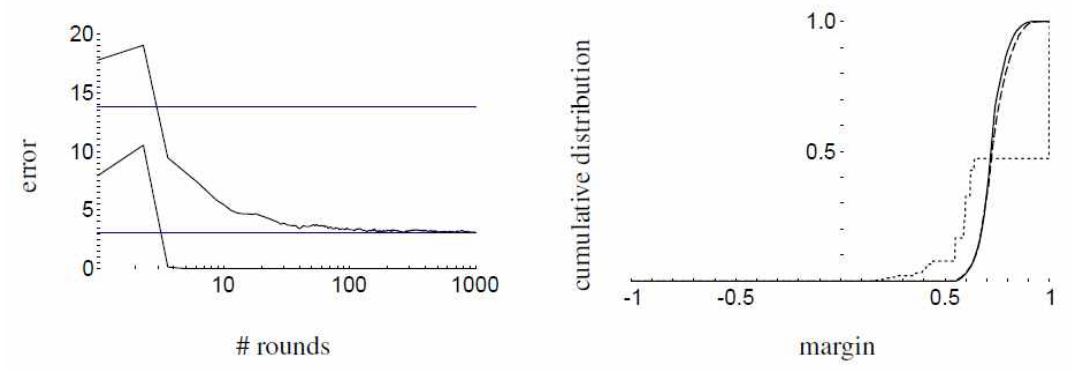
11) 본 연구에서도 선행연구와 마찬가지로 머신러닝을 이용한 분류모델과 유사한 전통적인 회귀분석 방법인 로짓모델을 비교하여 분석하기로 한다. 비교 분석 결과는 연구결과의 K-fold를 통한 모델 성능 비교 부분에서 다룬다.

제시하는 몇 가지 분류 모델에 대한 학습과 검정을 진행하여 가장 적합한 분류모델 선택과 지방소멸위험지역과 해당 지역의 특정 분야 취약성 등을 도출해본다.

1) 부스팅

부스팅은 앙상블 알고리즘 중(ensemble algorithm) 하나로 학습 알고리즘(learning algorithm)의 정확도 향상을 높이기 위한 가장 널리 알려진 방법으로, 라운드(rounds)의 거듭 시행을 통해 모델의 오류(error)를 줄이는 과정으로 볼 수 있다. 이러한 과정은 아래의 〈그림 2〉를 통해 살펴볼 수 있다.

〈그림 2〉 부스팅 과정에서 Error curves and Margin distribution 시각화



위 〈그림 2〉에서 살펴볼 수 있듯이 좌측 그래프를 통해 훈련, 검정 오류(training, test error)가 라운드를 거듭하면 낮아지는 것을 볼 수 있으며, 우측 그래프에서 트레이닝 예제에서 마진의 누적분포는 100~1000 사이에서 급격히 늘어나는 것을 볼 수 있다. 이러한 과정을 통해 어떠한 모델이 적합한지 찾을 수 있으며, 약한 예측 모형을 결합하여 강한 예측 모형을 만드는 알고리즘을 형성하는 것 또한 가능하다(Freund et al., 1999). 본 연구에서는 이러한 앙상블 부스팅 방법을 통해 가장 적합한 모형을 선정하여 제시한다.

2) Gradient Boosting Machine(GBM)과 XGBoost[12)]

앞서 서술한 바와 같이 부스팅 방법을 통해 모델의 성능을 향상시킬 수 있고, 행렬(table) 형태의 데이터에서 높은 예측 성능을 기대할 수 있기 때문에 많은 데이터 과학자들이 해당 알고리즘을 활용하여 분류 모델의 정확도를 높이고자 한다. Gradient Boosting Machine(GBM[13)])은 회귀분석 또는 분류분석을 수행할 수 있는 모델로, 예측모형의 앙상블 방법론 중 부스팅 계열에 속하는 알고

12) 해당 항의 내용은 Chen, T., & Guestrin, C. (2016)의 논문에 근거함.

13) 본 연구에서는 분류 모델이기 때문에 Gradient Boosting Classifier로 향후 서술하며, GBC를 약자로 한다.

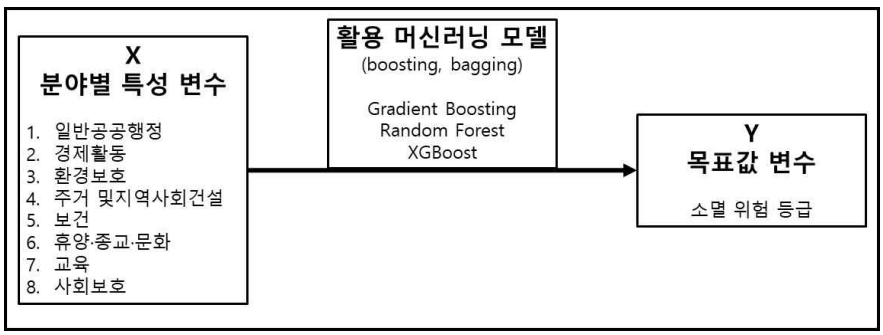
리즘으로 볼 수 있다. 그 중 XGBoost는 가장 효과적이고 많이 쓰이는 머신러닝 방법 중 하나인 트리 기반 부스팅(Tree boosting) 방법의 일종으로 현재 데이터 과학자들이 좋은 결과를 내는데 가장 많이 쓰는 머신러닝 방법이다.

이 방법은 적은 데이터에 대해 새로운 희소성 인식 알고리즘(Sparsity-Aware Algorithm)을 제안하고, 근사적 트리 학습(Approximate Tree Learning)을 위한 Weighted Quantile Sketch를 제안하는 방식이다. 측정 가능한(Scalable) 트리 부스팅 시스템을 구축하기 위해 캐시(Cache) 액세스 패턴, 데이터 축소(압축; Compression) 및 세분화(Sharding) 에 대한 통찰을 제시함으로써 기존 트리 기반 부스팅 방법보다 적은 자원을 사용하여 수십억 개 이상의 예제들을 다루어 측정하는 방법으로 볼 수 있다. 해당 방법의 장점으로 pGBRT, Spark MLLib, scikit-learn, R GBM 등의 기존 분류 모델보다 적은 자원을 가지고도 빠른 연산 속도와 함께 적합한 분류 학습도 가능하다고 볼 수 있다 (Chen & Guestrin, 2016).

3) Random Forest

앞서 서술한 부스팅 방법은 주요 앙상블 알고리즘 중 하나로 모델의 정교화를 도출할 수 있는 방법이다. 하지만 부스팅 방법은 과적합(overfitting)의 가능성이 있기 때문에 앙상블 알고리즘의 다른 방법인 배깅(bagging) 중 하나로 Random Forest 방법을 선택하여 모델에 대한 학습 성능을 검정할 수 있다. 배깅이란 bootstrap aggregating의 줄임말로 통계적 분류와 회귀 분석에서 사용되는 기계 학습 알고리즘의 안정성과 정확도를 향상시키기 위해 고안된 일종의 앙상블 학습법의 메타 알고리즘으로 볼 수 있다. 이에 속하는 Random Forest 방법은 의사결정나무(Decision Tree)를 다수 구성 후 앙상블(ensemble)하여 학습 성능을 높이는 방법으로 볼 수 있다. 상기한 분류 모델을 적용한 분석을 요약하여 도식화하면 다음 〈그림 3〉 과 같다.

〈그림 3〉 분류 모델 선정과 연구모형 도식화



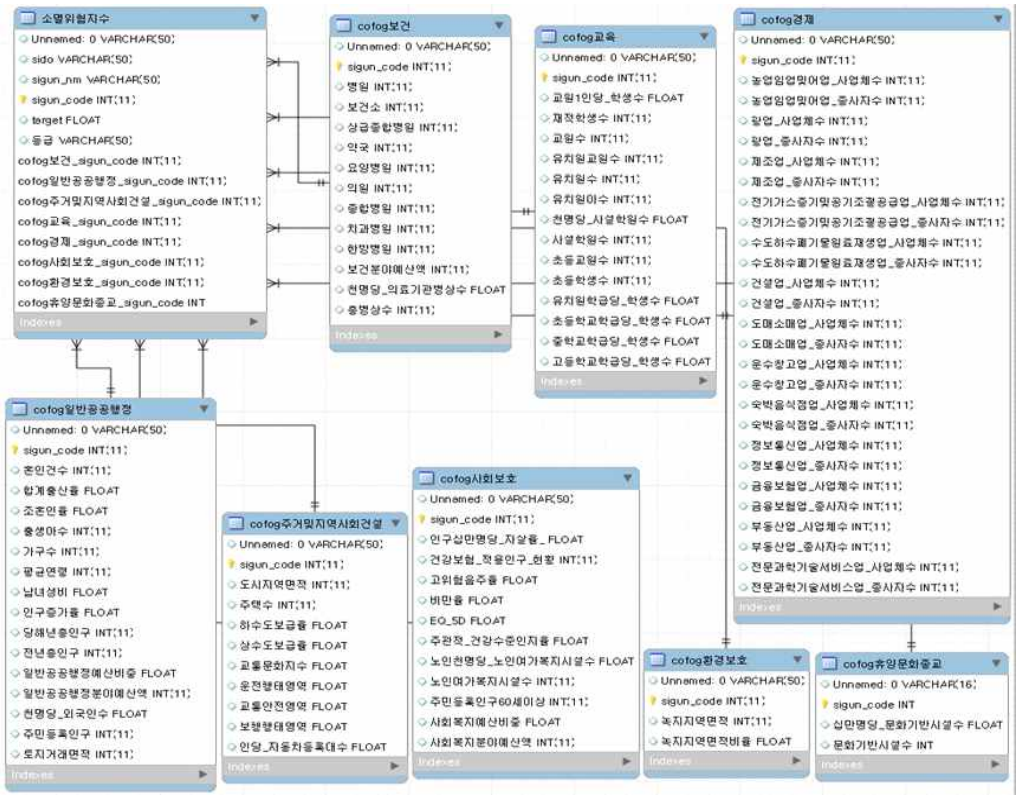
〈그림 3〉과 같이 본 연구에서는 앞서 서술한 방법을 사용하여 분류모델의 정확도를 향상시키고 가장 모델의 적합도가 높은 모델을 선정하여 학습 결과를 도출한다. 이를 통해 소멸위험에 영향을

미치는 특정 요인들을 도출하는 논의를 통해 향후 소멸위험지역을 예측하고, 요인을 제거하는 정책적 함의를 제시함으로써 해당 소멸위험지방자치단체가 소멸위험에서 벗어날 수 있는 방법에 대한 고찰을 진행해보기로 한다.

Ⅳ. 연구결과

본 연구에서는 정형 데이터를 활용한 머신러닝 분류모델을 구축하여 분류 결과를 도출한 후 소멸위험 지방자치단체로 분류된 지방자치단체를 도출하며, 이러한 지방소멸에 미치는 영향요인을 파악한다. 먼저, 정형 데이터를 활용한 분류모델에 사용할 데이터는 MariaDB 데이터베이스 연동을 통해 모든 데이터를 데이터베이스화하여 진행하였다. 앞서 연구설계 절에서 서술한 바와 같이 국가통계포털을 통해 자료를 수집한 후 소멸위험지수를 계산하고, 변수들을 COFOG 분류에 맞게 분류하여 데이터베이스화를 진행한다. 데이터베이스 구축은 2015~2018년 자료가 모두 있는 COFOG 분류 분야에서 진행하며, 이러한 과정을 거쳐 구축된 데이터베이스는 아래와 〈그림 4〉와 같다.

〈그림 4〉 데이터베이스 도식화



위 〈그림 4〉와 같이 소멸위험지수를 비롯하여, 8개 분야를 데이터베이스화하였으며, 2015~2016년 자료가 누락된 공공질서 및 안전 데이터는 자료 수집의 한계로 데이터베이스화하지 못하였다. 또한 앞서 변수 설명에서 제시한 바와 같이 국방 분야는 지방자치단체 논의와 합치되지 않기 때문에 제외하였다. 위와 같이 데이터베이스를 구축한 후 SQL 방법을 이용하여 데이터베이스와 연동하여 pandas 라이브러리로 데이터 구조화(frame)와 분석을 진행한다. 2015~2017년까지의 자료가 훈련자료(train set)로 활용되며, 2018년 자료가 시험자료(test set)로 구성된다. 훈련자료(train set)와 시험자료(test set)의 자료를 살펴보면 훈련자료(train set)의 행은 3개 년도를 반영한 684개로 구성되고, 시험자료(test set)의 행은 228개로 구성된다. 두 set 모두 93개의 열을 가지고 있으며, 해당 열이 변수로 반영된다. 해당 데이터프레임의 기초통계량을 확인한 화면을 제시하면 다음 〈그림 5〉 과 같다.

〈그림 5〉 기초통계량(train set, test set)

```
In [4]:
#기초 통계량 확인
train.describe()
```

Out [4]:

	sigun_code	병원	보건소	상급종합병원	약국	요양병원	의원	종합병원	치과병원	한방병원	...	숙박음식점업_종사자수	정보통신사업
count	684.000000	684.000000	684.000000	684.000000	684.000000	684.000000	684.000000	684.000000	684.000000	684.000000	...	684.000000	684.000
mean	37610.293860	6.543860	1.059942	0.187135	94.220760	6.328947	132.628655	1.305556	0.975146	1.248538	...	7013.421053	146.824
std	11671.293746	7.334477	0.617631	0.481041	89.561431	6.478367	161.287598	1.580972	1.852772	3.512735	...	8008.737058	413.441
min	11110.000000	0.000000	0.000000	0.000000	2.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	158.000000	3.000
25%	28597.500000	2.000000	1.000000	0.000000	24.000000	2.000000	24.000000	0.000000	0.000000	0.000000	...	1307.000000	14.000
50%	42725.000000	4.000000	1.000000	0.000000	61.000000	4.000000	74.000000	1.000000	0.000000	0.000000	...	4312.000000	30.000
75%	46722.500000	9.000000	1.000000	0.000000	142.000000	9.000000	186.500000	2.000000	1.000000	1.000000	...	9796.000000	96.250
max	50130.000000	44.000000	6.000000	3.000000	471.000000	38.000000	1518.000000	11.000000	17.000000	40.000000	...	58385.000000	4290.000

8 rows × 93 columns

```
In [5]:
#기초 통계량 확인
test.describe()
```

Out [5]:

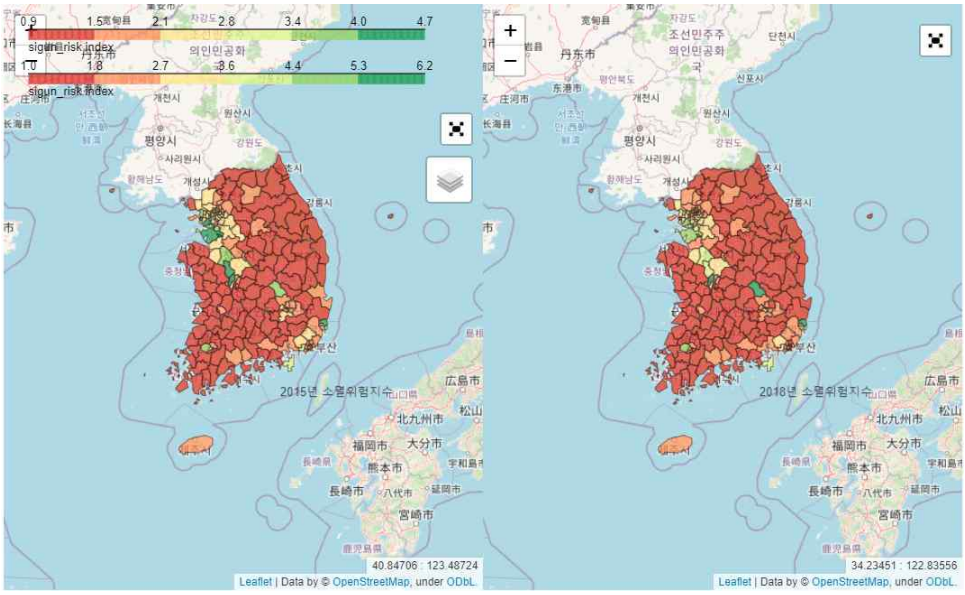
	sigun_code	병원	보건소	상급종합병원	약국	요양병원	의원	종합병원	치과병원	한방병원	...	숙박음식점업_종사자수	정보통신사업
count	228.000000	228.000000	228.000000	228.000000	228.000000	228.000000	228.000000	228.000000	228.000000	228.000000	...	228.000000	228.000
mean	37610.293860	6.425439	1.057018	0.184211	96.850877	6.842105	139.114035	1.364035	1.039474	1.346491	...	7613.754386	149.381
std	11688.419644	7.163211	0.616437	0.479827	92.745565	7.034940	170.962435	1.626847	1.868643	3.520630	...	8580.580267	401.447
min	11110.000000	0.000000	0.000000	0.000000	2.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	233.000000	2.000
25%	28597.500000	1.750000	1.000000	0.000000	23.000000	2.000000	25.750000	0.000000	0.000000	0.000000	...	1347.000000	14.000
50%	42725.000000	4.000000	1.000000	0.000000	60.500000	5.000000	76.500000	1.000000	0.000000	0.000000	...	4982.500000	33.000
75%	46722.500000	9.000000	1.000000	0.000000	146.000000	9.000000	197.500000	2.000000	1.000000	1.000000	...	10456.250000	108.500
max	50130.000000	39.000000	6.000000	3.000000	469.000000	40.000000	1588.000000	11.000000	16.000000	32.000000	...	55341.000000	3974.000

8 rows × 93 columns

데이터프레임에서 살펴볼 수 있듯이 전국 시군구 기준 지방자치단체의 2015~2017년의 기간 동안 수집된 COFOG 분야 변수가 훈련자료(train set)로, 2018년 기간에 수집된 COFOG 분야 변수가 시험자료(test set)로 활용된다. 이러한 데이터프레임을 활용하여 향후에 활용할 독립변수(feature)와 종속변수(target)를 로드(load)하며, 종속변수의 경우 지방소멸위험지수[14] 도출 후 연속형 변수인 지방소멸위험지수의 범주화하여 분석한다. 이러한 과정은 파이썬(python) 프로그램의 json, folium, matplotlib 등의 패키지를 활용하며, 해당 분석에서 지방소멸지수를 먼저 시각화하면 아래 〈그림 6〉 과 같다.

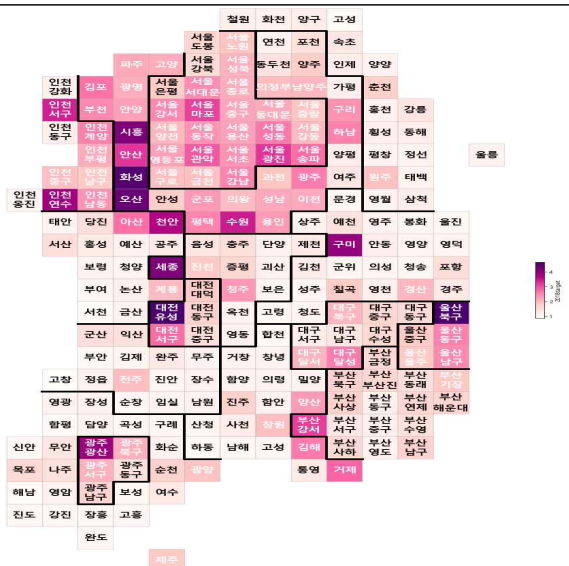
14) 앞서 서술한 인구이동을 반영하지 못하는 기존 소멸위험지수의 한계에 근거하여 해당 부분에서는 개선된 소멸위험지수를 사용하여 분석하였다.

〈그림 6〉 소멸위험지수 시각화(2015년, 2018년)



위 〈그림 6〉의 좌측은 2015년의 소멸위험지수를 시각화한 것이며, 우측은 2018년의 소멸위험지수를 시각화하여 나타낸 것이다. 소멸위험지수가 가장 높은 지역은 A등급으로 지도에서 진한 붉은색으로 표시된 곳으로 볼 수 있고, 해당 지역의 소멸위험지수는 약 1.5 미만으로 볼 수 있다. 〈그림 6〉에서 제시한 등급화에 라벨링을 진행하여 제시하면 아래 〈그림 7〉과 같이 볼 수 있다.

〈그림 7〉 소멸위험등급(target value) 시각화(2018년 기준)

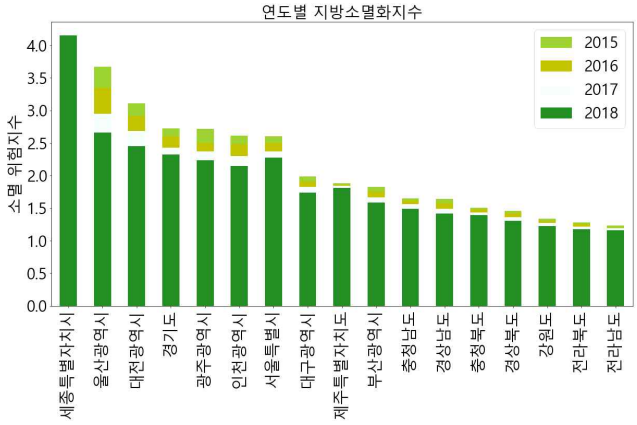


〈그림 7〉에서 볼 수 있듯이 색이 옅은 강원도, 경상남도, 경상북도, 인천광역시, 전라남도, 전라북도, 충청남도, 충청북도 등의 시도 내 시군구 지방자치단체에서 소멸위험을 확인할 수 있다.[15] 이렇게 예측된 연도별 지방소멸위험지수를 종속변수(target value)로 훈련자료(train set)와 시험자료(test set)를 구성하여 회귀분석 및 머신러닝 분석을 진행한다. 이를 위해 자료에 대한 탐색적 자료 분석(EDA: Exploratory Data Analysis)을 진행하면 다음과 같다.

먼저 훈련자료(train set)의 결측치를 확인하여 제거하기 위해 결측치가 존재하는 변수를 탐색·시각화하고 해당 변수에 존재하는 결측치를 각 변수의 시도 수준에서의 평균치로 해당 시군구 결측치를 대체하여 결측치를 제거하거나 해당 시군구에 적절한 값으로 변경하였다. 결측치 확인 결과, 보행행태영역, 교통문화지수, 운전행태영역, 녹지지역면적, 녹지지역면적비율 등의 변수에서 결측치가 관측되었기 때문에 해당 결측치를 해당 변수(column)의 행정구역(시도) 평균을 기준으로 시군구 결측치를 대체하여 반영하였다. 이 과정에서 경상남도 함양군의 2017년 제조업종사자수는 중분류별 제조업종사자수인 943명으로 대체하였고, 인천광역시 옹진군의 녹지면적은 통계조사 시 누락된 것으로 판단하여 해당 시도의 평균으로 대체하였다. 이 과정 후 훈련자료(train set)는 결측치가 없는 것으로 확인되었다.

시험자료(test set)의 경우 같은 방법을 통해 변수의 결측치를 확인하였으며, 녹지지역면적비율과 녹지지역면적[16]에서 결측치가 관측되어 해당 결측치를 마찬가지로 해당변수를 시도 수준에서의 평균값으로 대체하여 분석에 활용하였다. 일례로, 인천광역시 옹진군의 경우, 자살자가 0명인 통계가 있어 자살률을 0으로 대체하였고, 녹지면적이 누락되어 있어 해당 시도수준에서의 평균값으로 대체 하였다. 앞서 제시한 과정에서 결측치 확인하여 데이터 누락에 대한 문제를 해결한 후 시군구 및 시도를 기준으로 지방소멸지수 연도별 증감에 대한 분석을 진행하였다. 시군구의 소멸지수는 최종적으로 제시하기로 하고, 시도의 지방소멸지수는 해당 내용은 다음 〈그림 8〉 과 같다.

〈그림 8〉 연도별 지방소멸지수 변화



15) 색이 옅을수록 소멸위험이 높다.

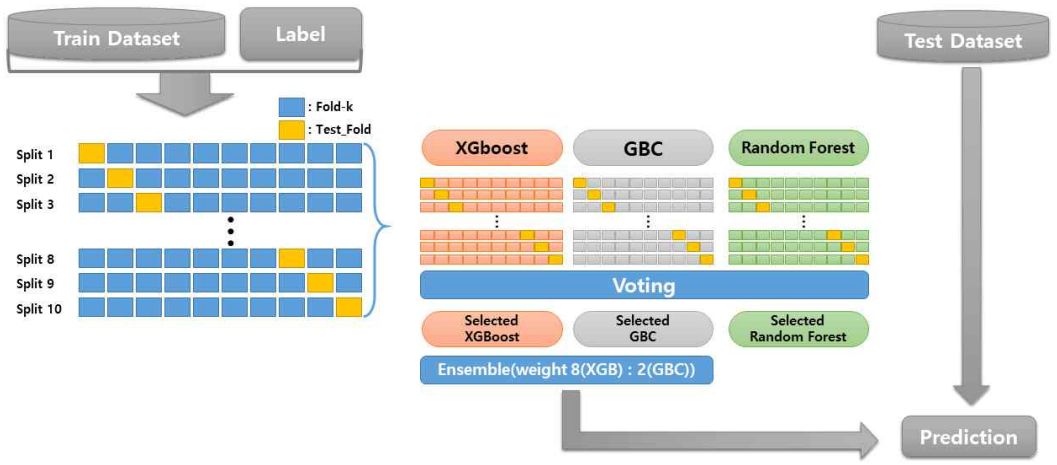
16) Missing Ratio: 녹지지역면적비율 = 0.438596, 녹지지역면적 = 0.438596

〈그림 8〉과 같이 지방소멸지수[17]의 연도별 변화를 분석하였을 때 지방소멸지수가 가장 높지만 감소폭이 큰 3개의 지역은 울산광역시, 대전광역시, 경기도로 볼 수 있고, 지방소멸지수가 가장 낮지만 감소폭이 안정적인 3개의 지역은 강원도, 전라북도, 전라남도로 볼 수 있다. 대전광역시와 울산광역시는 감소폭이 크다고 파악할 수 있으나 해당 도시는 경상도, 전라도, 충청도와 비교했을 때 소멸위험이 낮은 수준이므로 크게 우려할 수준은 되지 않는다고 볼 수 있다.

다만, 울산광역시, 대전광역시와 같이 광역시 수준에서도 지방소멸지수가 높아질 수 있음을 확인할 수 있다는 것은 여성인구 감소와 지방 간 인구이동이 적어졌다는 반증이기 때문에 시도 수준에서도 지방소멸에 대한 논의가 필요함을 역설할 수 있다. 이러한 지방소멸지수를 A~D로 나누어 각 등급을 라벨 인코더를 통해 각 시군구의 등급을 1~4로 바꾸어 모델링에 적합한 벡터로 전환하였다. 이때 A등급은 1, D등급은 4이다.

더하여 독립변수로 반영하는 요인들에 대한 다중공선성을 확인한 결과 변수 간 공선성이 존재함[18]을 탐색적으로 확인하였다. 하지만 본 연구에서는 분류모델에 영향을 미치는 변수들을 모두 반영하여 분류 모델의 성능을 향상시키고, 지방소멸위험지역을 정확히 분류해 내는 모델을 생성해내는 것이 연구의 목적이기 때문에 분류 모델에는 앞서 〈표 1〉에서 구성한 변수를 모두 반영한 모델을 구성하여 분석을 진행하였다. 아래 〈그림 9〉와 같은 과정을 통해 모델 구성 과정을 도식화 할 수 있다.

〈그림 9〉 모델 구성 과정의 도식화



본 연구의 연구 진행은 〈그림 9〉와 같이 도식화 할 수 있다. 먼저 국가통계포털에서 수집한 자료를 이용하여 변수를 구성하고, 훈련자료(train set)와 시험자료(test set)를 분류한 후 Stratified K-fold를 이용하여 자료 유효성을 높이며, XGBoost, GBC, RF 등의 모델을 구축하여 분석하고,

17) 소멸위험지수는 높을수록 소멸위험이 적다.
18) VIF 검정 10이상의 값

voting을 진행하여 각 모델 중 선택된 모델을 앙상블 모형에 반영하여 시험자료(test set) 결과를 예측한다. 이러한 일련의 과정을 후술하기로 한다.

먼저 K-폴드(이하 K-fold) 교차검증 방법은 본 연구에서 다루는 자료의 구성이 적은 약점을 개선하고 검증 성능을 높이고자 적용하였다. 자세히 설명하면 데이터의 수가 적은 경우에 해당 데이터 중 일부 검증 데이터의 수도 적기 때문에 검증 성능의 신뢰도가 떨어진다. 하지만 해당 데이터 내에서 검증 데이터의 수를 증가시키면 훈련용 데이터의 수가 적어지므로 정상적인 학습이 되지 않는다. 이러한 딜레마를 해결하기 위한 검증 방법이 K-fold 교차검증 방법으로 볼 수 있다. 본 연구에서는 K-fold 교차검증 방법을 중에서 불균형한 분포도를 가진 레이블 데이터 집합을 위한 방식인 Stratified K-fold를 활용하여 진행한다. 해당 결과는 아래 〈표 2〉과 같다.

〈표 2〉 Stratified K-fold 결과

〈다중공선성 있는 변수를 제거〉

	CrossValMeans	CrossValerrors
XGB	0.858396	0.032603
KNeighbors	0.526522	0.031844
RF	0.869768	0.031219
GBC	0.856821	0.025181
DT	0.763014	0.031832
Logist	0.430790	0.105127
SVC	0.291805	0.101293

〈다중공선성 있는 변수 제거하지 않음〉

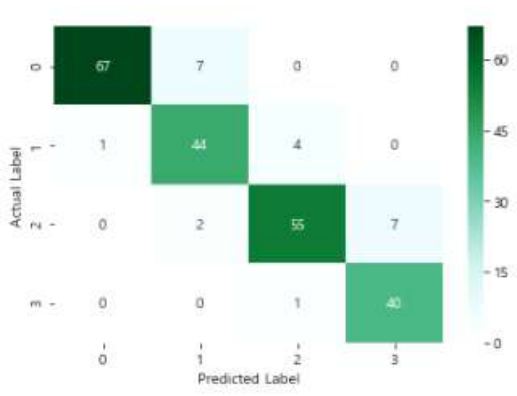
	CrossValMeans	CrossValerrors
XGB	0.888361	0.036470
KNeighbors	0.522065	0.078578
RF	0.887933	0.029035
GBC	0.885359	0.034012
DT	0.813907	0.047284
Logist	0.282241	0.067356
SVC	0.359241	0.089164

〈표 2〉는 데이터 세트를 10개의 그룹으로 분할한 후, 10개의 모델에 대해서 기본 모델을 구축한 뒤 파라미터 튜닝(RandomizedSearchCV)을 통해서 변수들을 최적화시켜 모델을 시각화하여 K-fold 분석을 진행한 결과로 이 결과 가장 성능이 좋은 3개의 모델(XGB, GBC, RFC)을 이용해서 앙상블을 진행한다. 해당 모델을 선택한 이유는 〈표 2〉에서 나타난 바와 같이 모델의 성능이 앞서 연구설계 부분에서 언급한 바와 같이 선택된 머신러닝 모델(XGB, RF, GBC)에서 다중공선성을 제거하지 않았을 때 모델의 성능이 더 높았으나, 로짓모델(logistic regression model: 〈표 2〉 내 Logist)의 경우 다중공선성의 문제로 모델의 성능이 저하되는 것을 확인할 수 있었기 때문이다. 선택된 GBC, RFC, XGB 모델에 대한 간단한 설명과 분석결과를 차례대로 제시하면 다음과 같다.

첫째, Gradient Boosting은 머신러닝 앙상블 기법 중 하나로 일련(sequential)의 약한 학습자(weak learner)들을 여러 개 결합하여 예측 혹은 분류 성능을 높이는 알고리즘인 부스팅(boosting)의 일종으로 볼 수 있다. 이는 일련(sequential)의 약한 학습자(weak learner)들의 잔차(residual)를 줄이는 방향으로 결합하여, Object Function과의 손실(loss)을 줄여나가는 개념체계에 근거하고 있으며, 여기서 정의되는 잔차(residual)는 Negative Gradient와 같은 의미를 지니게 되므로 Gradient

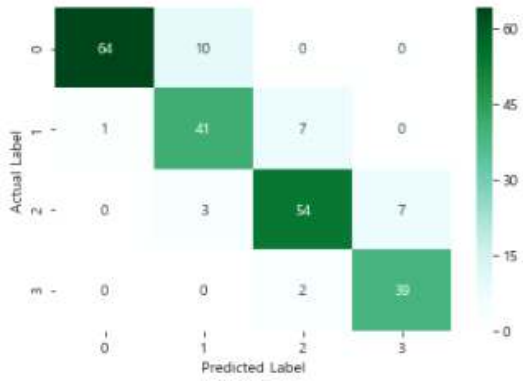
Boosting으로 정의하고 있다. 이러한 특성에 따라 해당 분석을 적용하고 진행하기 위해 먼저 파라미터 튜닝을 진행하고 최적화된 파라미터를 적용하여 F1 score, validation score를 도출하였다. F1 score의 경우, 데이터가 불균형한 구조일 때 모델의 성능을 정확하게 평가할 수 있기 때문에 지방소멸위험의 경우 소멸위험이 적은 지방자치단체와 소멸위험이 높은 지방 자치단체가 불균형하므로 F1 score를 채택한다. 해당 결과는 〈표 3〉로 제시한다.

〈표 3〉 Gradient Boosting Classifier 결과

<table>
<tr><th>RandomizedSearchCV 파라미터 최적화를 이용한
stratifed K-fold 결과</th><th>모델을 통해 시험자료를 예측하고
F1 score 측정한 결과</th></tr>
<tr><td rowspan="2">{'subsample': 0.95,
'n_estimators': 608,
'min_samples_split': 0.075,
'min_samples_leaf': 60,
'max_features': 'sqrt',
'max_depth': 15,
'loss': 'deviance',
'learning_rate': 0.025,
'criterion': 'friedman_mse'}</td><td>F1 Cross_validate 0.8915582178411123
F1 Macro: 0.8905957255158861
F1 Micro: 0.8947368421052632
F1 Weighted: 0.8956295791887187</td></tr>
<tr><td> 
</td></tr>
<tr><th>최적화된 파라미터를 반영한 GBC모델</th><th>voting 결과</th></tr>
<tr><td>GradientBoostingClassifier(learning_rate=0.2,
max_depth=30,
max_features='log2', min_samples_leaf=15,
min_samples_split=0.1,
n_estimators=773)</td><td>Validation score of a single GBC Classifier: 0.8828
Validation score of a VotingClassifier on 3 GBC with hard voting strategy: 0.8871
Validation score of a VotingClassifier on 3 GBC with soft voting strategy: 0.8871</td></tr>
</table>

〈표 3〉 의 결과에서 볼 수 있듯이, 해당 모델의 F1 score는 0.8915, Validation score of a single GBC Classifier: 0.8828로 판별되었다. 다음으로, Random Forest 모델 또한 마찬가지로 파라미터 최적화는 K-fold를 이용하여 진행하고, 최적화된 파라미터를 이용하여 Random Forest 모델을 만들어 같은 방식으로 분석을 진행한다. 해당 모델의 결과는 아래 〈표 4〉과 같다.

〈표 4〉 Random Forest 결과

<table>
<tr><th>RandomizedSearchCV 파라미터 최적화를 이용한
stratifed K-fold 결과</th><th>모델을 통해 시험자료를 예측하고
F1 score 측정한 결과</th></tr>
<tr><td rowspan="2">{'bootstrap': False,
'max_depth': 10,
'max_features': 7,
'min_samples_leaf': 1,
'min_samples_split': 3,
'n_estimators': 100}</td><td>F1 Cross_validate 0.8938804577096446
F1 Macro: 0.8659813097830549
F1 Micro: 0.868421052631579
F1 Weighted: 0.8699009261842074</td></tr>
<tr><td> 
</td></tr>
<tr><th>최적화된 파라미터를 활용한 RF모델</th><th>voting 결과</th></tr>
<tr><td>RandomForestClassifier(bootstrap=False,
max_depth=10, max_features=7,
min_samples_split=3)</td><td>Validation score of a singl rf Classifier: 0.8698
Validation score of a VotingClassifier on 3 rf with soft voting strategy: 0.8740
Validation score of a VotingClassifier on 3 rf with hard voting strategy: 0.8698</td></tr>
</table>

해당 모델에서는 F1 score은 0.8938, Validation score of a single RF Classifier: 0.8698로 판별되었다. GBC 모델과 비교해보았을 때 성능이 개선되지 않는 것을 알 수 있다. 마지막으로 XGB 모델을 같은 방법으로 파라미터 튜닝 후 XGBoost 모델을 구성하여 분석에 활용한다. GBM은 잔차(residual)를 줄이는 방향으로 일련(sequential)의 약한 학습자(weak learner)를 결합하기 때문에 성능측면에서 탁월하다고 할 수 있으나, 해당 훈련자료(train set)에 잔차를 계속 줄여나가기 때문에 과적합(overfitting)되기 쉽다는 문제점이 있다. 이를 해결하기 위해 XGBoost를 적용할 수 있는데 XGBoost는 GBM에 정규항(regularization term)을 추가한 알고리즘이다. 또한 다양한 손실함수(loss function)를 지원해 임무(task)에 따른 유연한 튜닝이 가능하다는 장점이 있다. 해당 모델의 결과는 아래 〈표 5〉 과 같이 나타낼 수 있다.

〈표 5〉 XGBoost 결과

<table>
<tr><th>RandomizedSearchCV 파라미터 최적화를 이용한
stratifed K-fold 결과</th><th>모델을 통해 시험자료를 예측하고
F1 score 측정한 결과</th></tr>
<tr><td rowspan="2">{'subsample': 0.88,
'reg_lambda': 0,
'num_class': 2,
'n_estimators': 1698,
'min_child_weight': 1,
'max_depth': 25,
'learning_rate': 0.1,
'gamma': 0.01,
'eta': 0.1,
'colsample_bytree': 0.3,
'colsample_bylevel': 0.9}</td><td>F1 Cross_validate 0.8986524393947791
F1 Macro: 0.8878868459626538
F1 Micro: 0.8947368421052632
F1 Weighted: 0.8956796442703215</td></tr>
<tr><td> 
</td></tr>
<tr><th>최적화된 파라미터를 XGBoost 모델</th><th>voting 결과</th></tr>
<tr><td>XGBClassifier(base_score=0.5, ooster='gbtree',
colsample_bylevel=0.9, colsample_bynode=1,
colsample_bytree=0.3, eta=0.1,
gamma=0.01,gpu_id=-1, importance_type='gain',
interaction_constraints='', learning_rate=0.1,
max_delta_step=0, max_depth=25,
min_child_weight=1, missing=nan,
monotone_constraints='()', n_estimators=1698,
n_jobs=0, num_class=2, num_parallel_tree=1,
objective='multi:softprob', random_state=0,
reg_alpha=0, reg_lambda=0,
scale_pos_weight=None, subsample=0.88,
tree_method='exact', validate_parameters=1,
verbosity=None)</td><td>Validation score of a single XGB Classifier: 0.8871
Validation score of a VotingClassifier on 3 XGB with
hard voting strategy: 0.8957
Validation score of a VotingClassifier on 3 XGB with
soft voting strategy: 0.8957</td></tr>
</table>

이 결과 실제 a등급을 a등급으로 제대로 예측한 데이터 67개로 확인할 수 있으며, F1 score 0.8986으로 가장 좋은 성능으로 측정되었고, confusion matrix 결과에서도 데이터가 잘 분류되었음을 확인할 수 있다. 이러한 결과를 반영하여 XGBoost 모델을 대상으로 성능 강화를 진행해보기로 한다. 성능 강화 이전에 이러한 일련의 분류과정에서 나타난 변수의 중요도와 순열 특성 중요도를 특정하면 다음 〈표 6〉와 같다.

〈표 6〉 변수 중요도 분석결과

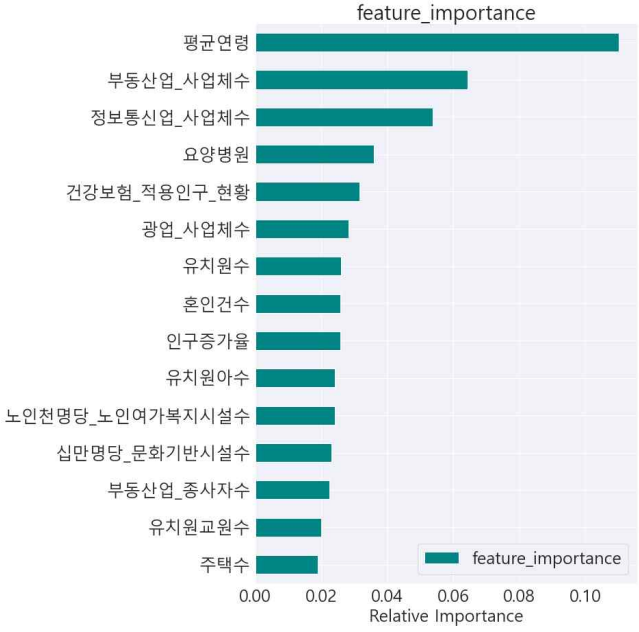
변수 중요도
(feature importance)

	feature_importance
평균연령	0.110163
부동산업_사업체수	0.064286
정보통신업_사업체수	0.053803
요양병원	0.035866
건강보험_적용인구_현황	0.031473
광업_사업체수	0.028266
유치원수	0.025903
혼인건수	0.025803
인구증가율	0.025798
유치원아수	0.024117
노인천명당_노인여가복지시설수	0.024081
십만명당_문화기반시설수	0.023125
부동산업_종사자수	0.022414
유치원교원수	0.019942
주택수	0.018830

순열 특성 중요도
(permutation importance)

Weight	Feature
0.3049 ± 0.0232	평균연령
0.0459 ± 0.0078	십만명당_문화기반시설수
0.0386 ± 0.0092	인구증가율
0.0205 ± 0.0018	노인천명당_노인여가복지시설수
0.0023 ± 0.0023	혼인건수
0.0015 ± 0.0000	광업_사업체수
0.0003 ± 0.0012	조혼인율
0 ± 0.0000	교통안전영역
0 ± 0.0000	주민등록인구60세이상
0 ± 0.0000	교원1인당_학생수
0 ± 0.0000	재적학생수
0 ± 0.0000	교원수
0 ± 0.0000	교통문화지수
0 ± 0.0000	운전행태영역

변수 중요도 시각화



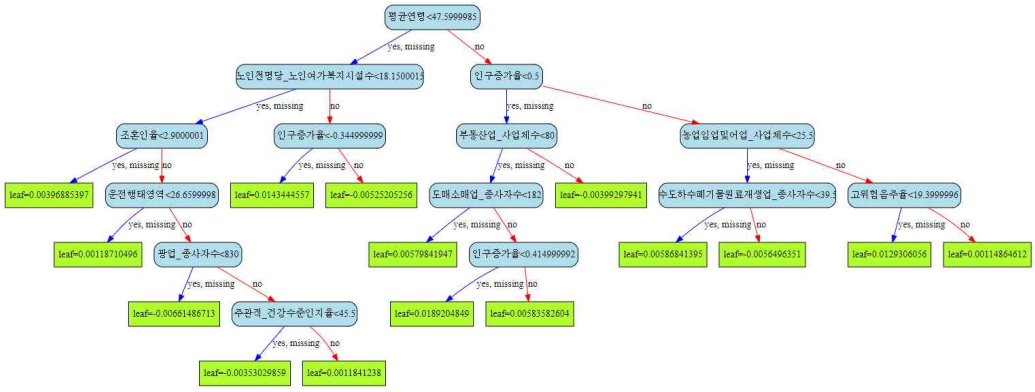
〈표 6〉 변수 중요도 분석 결과는 본 연구에서 사용한 트리 기반 모형에서 특정 변수들이 트리를 분할시키는데 얼마나 기여했는지에 대한 정도를 나타낸다. 트리를 분할하는 원리는 불순도를 기

준으로 분할하게 되며, 주어진 샘플을 잘 분류시킬수록 불순도는 낮아지게 된다. 본 연구에서 사용되는 모형에서는 불순도를 측정하는 지표로써 Gini 계수를 이용하며, 아래와 같은 수식을 통해 산출한다(Ture et al, 2009).

$$gini(T) = 1 - \sum_{i=1}^{n} p^{2_i}$$

수식에서 p는 범주형 변수의 경우 범주 수를 나타내고, 연속형 변수의 경우 해당변수를 오름차순으로 정렬했을 때, 종속변수의 범주가 바뀌는 수 나타낸다. 하지만 Gini 계수만으로는 노드의 불순도만을 측정할 뿐, 어떤 변수가 모형에 중요한 영향을 미치는지는 파악할 수 없다(Shmueli et al, 2019). 따라서 원래 노드의 불순도에서 특정 변수를 선택·분류하여 산출되는 불순도의 값을 빼주었을 때 가장 낮은 변수를 선택함으로써 분할 기준으로 사용하게 되며, 이때 변수별로 계산된 값을 변수 중요도로 정의한다. 이러한 일련의 과정은 다음 〈그림 11〉에서 제시할 수 있다.

〈그림 11〉 변수 중요도 도출과정에서의 트리기반 모형 시각화



하지만 이러한 변수 중요도라는 지표 역시 노드가 분기할 때마다의 지니계수만을 고려하여 중요도를 부여하기 때문에 과적합 문제에 대한 고려를 하지 못한다는 한계가 존재한다. 이러한 이유로 본 연구에서는 순열 특성 중요도라는 지표를 추가하여 변수 중요도의 타당성을 확보하고자 하였다. 순열 특성 중요도는 특정 변수가 모형에서 중요한 역할을 하고 있다면 해당 변수를 모형에서 제거했을 때 성능이 크게 떨어질 것이라는 개념을 적용한 지표로써, 모형에 사용된 변수를 변화시켜가며 성능을 계산한 값을 나타낸다. 여기서 가중치(weight)로 표시된 성능 감소량은 더 큰 양수의 값을 가질수록 모델에 큰 영향을 끼친다는 의미로 해석되어 변수 중요도의 타당성을 확보하는 장치로 사용할 수 있다(Shmueli et al, 2019).

결과적으로 〈표 6〉을 살펴보면 평균연령, 부동산업/정보통신업 사업체 수, 요양병원, 건강보험

적용인구, 광업 사업체 수, 유치원 수, 혼인 건 수, 인구증가율, 유치원아 수, 노인여가 복지시설 수, 문화기반시설 수, 부동산업 종사자 수, 유치원 교원 수, 주택 수 등이 중요한 변수로 나타났다. 이러한 변수들은 경제적인 특성을 갖는 변수들과 의료시설, 문화시설과 같은 지역의 매력도(amenity)를 높일 수 있는 변수 혹은 소멸지수의 수식과 연관이 있거나 혼인 건 수과 같은 인구 사회학적 변수 등으로 볼 수 있는데 이는 지방소멸지수에 직접적 영향을 미친다고 볼 수 있다. 따라서 향후 소멸위험이 높은 지방자치단체들은 긍정적 요인인 지역의 매력도를 높일 수 있는 의료시설, 문화시설과 같은 다양한 긍정적 수단에 대한 확보가 필요하며, 경제적인 기반 또한 확보 및 관리가 필요하다는 결론을 1차적으로 도출할 수 있다.

더하여 XGBoost 모델의 성능이 가장 좋음을 확인하였기 때문에 XGBoost에서 RandomizedSearchCV를 사용하여 n_estimators, max_depth, num_class 등 최적의 파라미터를 한차례 더 검색하여 XGBoost 모델의 성능을 향상시켜보기로 한다. XGBoost voting을 실시하여 클래스 불균형 문제를 해결하여 모델의 성능을 높이고자 하였다. XGBoost voting은 분류 가중 균형화된 y 값을 사용하여 입력 데이터의 클래스 빈도에 반비례를 통해 가중치를 자동으로 조정하고, 다중 출력의 경우 y의 각 열의 가중치가 곱해진다. 클래스 가중치(class weights)는 클래스 불균형을 수정하기 위해 가중치를 적절히 변경한다. 결국 voting은 서로 다른 머신러닝 알고리즘으로 여러 개의 분류기를 생성하고, 투표를 통해 최종 예측 결과를 결정하는 방식으로 볼 수 있다.

이 voting 결과 F1 score는 89.6%까지 상승하였으며, 228개 지방자치단체 중 소멸위험이 큰 지역을 소멸위험이 크다고 예측한 경우가 67개 지역, 소멸위험이 적은 지역을 소멸 위험이 적다고 예측한 경우가 37개 지역이다. 마지막으로, 앙상블을 통해 여러 예측 모형을 만들어 낸 후 최종 예측 모형을 정하고자 하였다. 모델 간 상관관계를 확인한 결과 XGB와 GBC 모델을 이용하여 앙상블을 진행하였다. 최종적으로 앙상블 모형의 성능은 XGBoost voting에서 일부 향상된 90%(0.899947)의 예측 성능을 보였으며, 소멸 위험이 큰 지역은 아래 〈표 7〉에서 확인할 수 있는 68개의 지방자치단체[19]이다.

〈표 7〉소멸 고위험 지방자치단체명

소멸위험이 높은 지방자치단체명(총 68개)[20]				
강원도 양양군	경상남도 합천군	경상북도 청송군	전라남도 장흥군	전라북도 진안군
강원도 영월군	경상북도 고령군	인천광역시 강화군	전라남도 진도군	충청남도 금산군
강원도 정선군	경상북도 군위군	인천광역시 옹진군	전라남도 함평군	충청남도 부여군
강원도태백시	경상북도 문경시	전라남도 강진군	전라남도 해남군	충청남도 서천군
강원도 평창군	경상북도 봉화군	전라남도 고흥군	전라북도 화순군	충청남도 예산군
경상남도 거창군	경상북도 상주시	전라남도 곡성군	전라북도 고창군	충청남도 청양군
경상남도 고성군	경상북도 성주군	전라남도 구례군	전라북도 김제시	충청남도 태안군

19) XGBoost 모형으로 소멸위험이 큰 지역을 소멸위험이 크다고 예측한 경우가 67인데, 〈표 7〉의 결과 68개인 이유는 XGBoost 단일 모델의 결과가 67개였으나, XGBoost와 GBC모델을 8:2비율로 해서 앙상블하여 예측을 할 때 성능이 90% 가장 높게 나왔고, 이 모델을 통해 예측한 소멸위험지방자치단체가 68곳이다.

경상남도 남해군	경상북도 영덕군	전라남도 담양군	전라북도 남원시	충청북도 괴산군
경상남도 산청군	경상북도 영양군	전라남도 보성군	전라북도 무주군	충청북도 단양군
경상남도 의령군	경상북도 영주시	전라남도 신안군	전라북도 부안군	충청북도 보은군
경상남도 창녕군	경상북도 울릉군	전라남도 영광군	전라북도 순창군	충청북도 영동군
경상남도 하동군	경상북도 울진군	전라북도 영암군	전라북도 임실군	충청북도 옥천군
경상남도 함안군	경상북도 의성군	전라남도 완도군	전라북도 장수군	
경상남도 함양군	경상북도 청도군	전라남도 장성군	전라북도 정읍시	

위 〈표 7〉에서 살펴볼 수 있듯이 강원도, 경상도, 인천광역시, 전라도, 충청도 등에서 소멸위험이 가장 높은 등급의 시군구가 존재함을 확인하였고, 해당 지방자치단체들은 공통적으로 인구사회학적으로 노인인구가 많고, 여성인구와 혼인건수가 적으며, 해당 지방자치단체로의 전입이 적다는 공통점이 있었다. 이러한 문제는 앞서 다양한 연구들에서도 진행되었기 때문에 차후 정책적 함의에서 상세한 고려를 진행해보기로 한다.

V. 결론

1. 연구결과 요약

본 연구에서는 정형 자료와 비정형 자료를 이용하여 각각 국가통계포털의 정형자료를 활용하여 머신러닝 분류모델의 구축, 분석을 진행하였다. 연구결과, 소멸위험이 높은 지방자치단체는 68개로 분석되었으며, 해당 분류를 진행한 분류모델은 최종적으로 XGBoost, GBC 모델을 활용한 앙상블(ensemble) 모형으로 90%의 성능을 확인할 수 있었다. 이 분석을 통해 지방자치단체의 소멸위험에 미치는 영향이 높다고 판단할 수 있는 변수들은 평균연령, 부동산업/정보통신업 사업체 수, 요양병원, 건강보험 적용인구, 광업 사업체 수, 유치원 수, 혼인 건 수, 인구증가율, 유치원아 수, 노인여가 복지시설 수, 문화기반시설 수, 부동산업 종사자 수, 유치원 교원 수, 주택 수이다.

이 결과에 근거하여 경제시설(사업체수, 취업률에 근거), 문화시설, 의료시설은 지역의 매력도를 높이는 역할을 하는 것으로 보인다. 취업률과 사업체는 경제적 변수로써 지역의 경제 기반을 의미할 수 있으며, 이러한 경제 기반이 잘 되어 있는 경우에는 소멸위험이 적어지는 것으로 판단된다. 더하여 문화시설은 특히 인간 삶의 질적 영향을 미치는 요인으로 판단할 수 있기 때문에 인간적이고 개인주의적이며 자존적인 현 20~30대들의 지역 선택의 경향에 이러한 변수가 충분히

20) 같은 머신러닝 방법을 사용하여 인구이동을 반영하지 않은 기존 소멸지수로 도출한 지방자치단체는 부록에서 제시하기로 한다. 기존소멸지수로 도출한 지방자치단체는 62개였으며, 인구이동 반영소멸지수로 도출된 결과와 6개 지방자치단체의 차이가 났고, 기존 소멸지수로 도출한 소멸위험 지방자치단체와 인구이동을 반영하여 소멸지수로 도출한 소멸위험 지방자치단체의 교집합을 제외하고 기존 소멸지수로 특수하게 도출된 지방자치단체는 강원도 횡성군과 경상북도 예천군이다.

영향을 미칠 수 있다고 본다. 의료시설의 확충도 마찬가지로 노령인구의 증가와 각종 질병에 노출된 현대인들에게 충분히 지역의 매력도를 높일 수 있는 요인으로 작용할 수 있다.

마지막으로 평균연령, 혼인건수, 인구증가율 등은 직접적으로 지방소멸위험지수를 도출하는 식에 포함되기도 하지만 지역의 활성화와 해당 지역 내에서 인구의 재생산이 일어나는지, 지역 외에서 인구의 유입이 되는지에 대한 지표이기 때문에 해당 지표 또한 지방소멸가능성에 지대한 영향이 있다고 판단된다. 이러한 면을 종합적으로 고려할 때 향후 지방자치단체들의 소멸가능성을 낮출 수 있도록 여러 요인에 대한 보완과 개선이 필요할 수 있다. 이는 다음 절의 정책적 함의에서 제시하기로 한다.

2. 정책적 함의

본 연구에서 머신러닝 분류 모델 결과를 제시하면서 소멸위험이 높은 지방자치단체는 인구사회학적 변수들이 저조하며, 경제, 문화, 의료 기반이 취약하여 소멸위험이 높아질 수 있음을 확인하였다. 이러한 결과를 통해 본 연구에서는 다음과 같은 정책적 함의를 다음과 같이 제시하고자 한다.

경제 기반 시설의 수도권 집중화 현상을 타파해야 한다. 본 연구의 GIS 분석 결과를 살펴보면 수도권, 부산, 울산, 대전 등과 같은 광역시는 상대적으로 소멸위험이 적음을 알 수 있다. 자세히 서술하면, 평균연령, 인구증가율, 혼인건수는 지방소멸지수에 영향을 미치는 주요 변수로 인구사회학적 변수에 해당한다. 해당 변수는 상관관계 분석 결과 지역의 경제적 기반인 사업체수, 취업률 등과 밀접한 연관이 있을 뿐만 아니라 의료시설, 문화시설 등과도 관련이 있는 것으로 나타났다. 지방 소멸은 결국 인구감소로 인해 해당 지방자치단체의 인구 공동화 현상이 나타나는 것으로 볼 수 있는데 이러한 공동화 현상은 기본적으로 경제 기반 시설의 부재에서 나타날 수 있다는 반증이 된다.

이러한 문제를 해결하기 위해 공공기관의 지방 이전은 올바른 정책 방안이라고 볼 수 있다. 다만, 최근 서울 지역에서 공공기관에 취업하여 지방으로 거주지를 옮기는 경우가 많아지고 있는데, 현재 지방대학 출신 인재들에게 한정적으로 주는 가산점을 해당 지방에서 고등학교를 졸업한 인재까지도 확대하여 우수한 대학교육을 받기 위해 귀경한 지방인재들을 다시 지방으로 돌아올 수 있도록 하는 취업 우대 정책의 확대도 필요하다고 보인다. 경제적 기반이 갖춰지고 이전 공공기관에 취업을 하는 인구가 많아지게 되면 해당 지역을 중심으로 상권이 발달하게 될 것이기 때문에 인구 유입을 증가시킬 수 있다.

이러한 경제활동인구의 증가는 소득 증가와 인구 증가로 이어지게 되기 때문에 자연히 세수(稅收)와 가구 수의 증가로 이어진다. 이는 지방소멸위험에서 벗어날 수 있는 적절한 방편이 될 수 있으며, 공공기관과 더불어 중소·대기업에는 세재 혜택이나 각종 세금을 탕감하여, 개별 업체가 아닌 특정한 산업 전체를 지방으로 유치하는 전략으로 규모의 경제를 확보하는 방안을 활용한다면 지방소멸을 막을 수 있는 경제적 기반을 구축할 수 있다. 지방 소멸 타파의 핵심은 경제적 기반의

확보 후 문화, 의료, 관광 시설의 확충까지의 이어지는 일련의 과정이라고 할 수 있다.

지방 매력도의 감소는 해당 지방으로 이주하려는 사람, 사업을 하려는 사람, 관광·방문하려는 사람들과의 거래비용을 증가시켜 지속적으로 지방소멸위험을 높일 가능성이 있다. 결국 이러한 지방매력도 감소는 지역에 대한 애착이나 긍정적인 인식을 갖는 사람들이 줄어든다는 것을 의미하기 때문에 이러한 문제를 해결할 필요가 있다. 최근 이러한 인식을 타파하기 위해 SNS(social network, social media)를 활용하는 지방자치단체가 많아지고 있는데 소멸위험 지방자치단체들 또한 지방자치단체의 긍정적 부분을 부각시키는 소셜 미디어 활동을 이어감과 동시에 지방자치단체들이 가지고 있는 다양한 주력 산업이나 문화·의료 시설 등에 대한 적극적인 홍보가 필요하다고 볼 수 있다.

더하여, 최근 지방의 의료 격차[21]와 문화 격차[22]가 서울 등의 수도권과 비교해 보았을 때 상당히 크다는 현실이 존재하기 때문에 이러한 부분에 있어서 공공의 역할이 필요할 것으로 보인다. 결국, 수도권과 지방의 지역 간 격차를 줄여 국민들이 자신들이 원하는 지방에서 적절한 경제활동과 여가생활, 의료서비스를 충분히 향유할 수 있다면 지방을 떠나는 사람들이 적어질 것이며, 새로이 지방으로 유입되는 인구도 많아질 가능성이 있다. 따라서 향후 지방소멸 극복을 위한 지방지원 정책도 이러한 부분에 초점을 맞춰 진행될 필요가 있다.

마지막으로 본 연구는 기존의 소멸지수를 개선하려는 노력과 동시에 머신러닝 방법을 활용하여 실증적으로 지방소멸예측과 그 영향요인을 추출하는 방식으로 지방소멸의 문제를 다루어 지방소멸의 위험이 높은 지역, 지방소멸에 영향을 미치는 요인을 실증적으로 분석·도출하려는 노력을 하여 정책적 함의를 제시하였다는 점에서 의의가 있을 것으로 보인다. 더하여 본 연구의 한계로는 머신러닝 모델을 활용하였으나, 데이터의 크기가 크지 않기 때문에 K-fold 방법론을 활용하여 유효성을 높였다는 점과 소멸위험지수에 인구이동을 반영한 이론적인 근거가 미미함에 대해서 한계가 있을 것으로 보인다.

참고문헌

Alin, A. (2010). Multicollinearity. Wiley Interdisciplinary Reviews: Computational Statistics, 2(3), 370-374.

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

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

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of

21) 지역 의료격차 이렇게 심하다니…강남 29.6명 vs 영양 107.8명, 매일경제, 2018.10.01. https://www.mk.co.kr/news/it/view/2018/10/612985/

22) 문화예술 분야 지역 불균형, 첫 지표로 확인해보니… 서울 활동지수 100일 때 충북 2.6 그쳐. 국민일보. 2015.08.18. http://news.kmib.co.kr/article/view.asp?arcid=0923204492

the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Choi, E. (2012) Urban amenities as determinants of selecting a logo type in Korea: the multinomial logit approach with the bootstrap sample. Quality & Quantity 46: 391-404.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In International workshop on multiple classifier systems (pp. 1-15). Springer, Berlin, Heidelberg.

Freund, Y. Schapire, R. and Abe, N. (1999). A short introduction to boosting. Journal-Japanese Society For Artificial Intelligence, 14(771-780), 1612.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

Gottlieb, P. D.(1994). Amenities as an economic development tool: Is there enough evidence?. Economic Development Quarterly 8(3), 270-285.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... and Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In Advances in neural information processing systems (pp. 3146-3154).

Logan, J. R. and Molotch, H. L.(1987). Urban Fortunes, Univ. of California Press.

McNulty, R. H., Jacobson, D. R. and Penne, R. L.: The Economics of Amenity: Community

Natekin, A., and Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in neurorobotics, 7, 21.

Shmueli, G., Patel, N. R., and Bruce, P. C. (2011). Data mining for business intelligence: Concepts, techniques, and applications in Microsoft Office Excel with XLMiner. John Wiley and Sons.

Shmueli, G., Bruce, P. C., Gedeck, P., and Patel, N. R. (2019). Data mining for business analytics: concepts, techniques and applications in Python. John Wiley & Sons.

Tang, L., Xiong, C., and Zhang, L. (2015). Decision tree method for modeling travel mode switching in a dynamic behavioral process. Transportation Planning and Technology, 38(8), 833-850.

Ture, M., Tokatli, F., and Kurt, I. (2009). Using Kaplan–Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4. 5 and ID3) in determining recurrence-free survival of breast cancer patients. Expert Systems with Applications, 36(2), 2017-2026.

Xian-Yu, J. C. (2011). Travel mode choice analysis using Support Vector Machines. In ICCTP 2011: Towards Sustainable Transportation Systems (pp. 360-371).

강동우. (2019). 지방소멸 위험과 지역고용의 상관관계 분석. 「노동리뷰」, 30-39.

강병수. (2014). 지역어메니티와 주거이전과의 관련성에 관한연구. 한국도시행정학회 학술발표대회 논문집, 329-351.

기영화. (2013). 지방정부 노인일자리사업의 사회적 자본 효과 탐색: 근거이론. 「지방정부연구」, 17(1), 265-295.

김동완. (2015). 지방 소멸의 시대, 생존전략으로서 균형발전. 월간 공공정책, 120, 28-30.
김서인·김동성·김종우. (2016). 국내 주요 10 대 기업에 대한 국민 감성 분석: 다범주 감성사전을 활용한 빅 데이터 접근법. 「지능정보연구」, 22(3), 45-69.
김성자·문하은·정성호. (2016). 정부 기능별분류 (COFOG) 의 체계적 고찰 및 활용. [BOK] 국민계정 리뷰, 2016(3).
김순은. (2017). 02 저출산·고령사회의 인구감소와 지방소멸로 인한 대응책. 「지방행정」, 66(759), 26-29.
_____. (2017). 일본의 지방창생정책. 「공공정책연구」, 33(2), 25-54.
김은혜·박배균. (2016). 2000 년대 이후, 일본의 국가 스케일 재편과 특구 전략. 공간과 사회, 10-43.
김인호·이경섭. (2020). 트리 기반 앙상블 방법을 활용한 자동 평가 모형 개발 및 평가: 서울특별시 주거용 아파트를 사례로. 「한국데이터정보과학회지」, 31(2), 375-389.
김현호·오은주. (2007). 어메니티를 활용한 지역발전 방안. 한국지방행정연구원 기본연구과제, 2007, 1-179.
남민지·이은지·신주현. (2015). 인스타그램 해시태그를 이용한 사용자 감정 분류 방법. 「멀티미디어학회논문지」, 18(11), 1391-1399.
박승현. (2017). '지방소멸'과 '지방창생': '재후'(災後) 의 관점으로 본 '마스다 보고서'. 일본비평 (Korean Journal of Japanese Studies).
소진광. (2004). 사회적 자본 형성을 통한 지방자치와 지역발전의 연계화 방안. 「지방행정연구」, 18(2), 67-91.
심재현. (2009). 주거환경과 어메니티: 아파트의 주거만족도를 중심으로. 「한국행정과 정책연구」, 7(1), 65-83.
안명준·배정한·주신하·신지훈·이동근. (2008). 농촌 어메니티 경관의 평가 체계 개발과 적용 -[2007 농촌 어메니티 100 선] 을 중심으로. 「농촌계획」, 14(2), 77-84.
원광희·채성주·설영훈. (2020). 지방소멸위험지수의 기준은 과연 적합한가?. 충북 FOCUS, 1-26.
이건웅. (2020). 지역소멸과 지역재생의 해결 방안 연구. 「글로벌문화콘텐츠」, 43, 125-143.
이기배. (2017). 일본의 인구감소 시대 지역발전정책의 체계 및 방향성에 관한 연구. 「도시행정학보」, 30(4), 81-104.
이상호. (2016). 지역고용동향 심층분석: 한국의 '지방소멸'에 관한 7 가지 분석. 지역고용동향 브리프, 한국고용정보원.
이영호·홍성연. (2019). 머신러닝을 활용한 개인의 교통수단 선택 예측모형 구축. 「한국데이터정보과학회지」, 30(5), 1011-1024.
이정환. (2017). 인구감소와 지속가능한 지방만들기-지방소멸 (地方消滅) 을 둘러싼 논점. 「일본공간」, 21, 194-223.
이호상. (2016). 지방소멸: 인구감소로 연쇄붕괴하는 도시와 지방의 생존전략: 마스다 히로야 지음, 김정환 옮김, 와이즈베리, 2015. 「인천학연구」, 24, 217-225.
이희태. (2012). 지방정부의 경쟁력 강화를 위한 사회적 자본 확충 전략: 부산광역시 해운대구를 중심으로. 「지방정부연구」, 16(3), 69-89.
임보영·이경수·마강래. (2018). 지방소멸과 저성장 시대의 국토공간전략: 일본의 사례를 중심으

로.「공간과 사회」, 64, 45-70.
임형백. (2012). 농촌 어메니티를 이용한 농촌활성화 정책 방향.「지방행정연구」, 26(3), 3-25.
정성호·홍창수. (2018). 강원 지역의 소멸 가능성에 관한 연구.「사회과학연구」, 57(1), 3-25.
_____. (2019). 강원도 인구변화와 지역소멸 위험.「사회과학연구」, 58(1), 3-22.
_____. (2019). 지방소멸론에 대한 비판적 검토.「지역사회학」, 20, 5-27.
진관훈. (2012). 지방정부의 사회적 자본 증대 방안 연구.「지방정부연구」, 16(3), 395-412.
최길수·정영윤·방정희(2013). 지역단위의 사회적자본 측정 및 관리방안에 관한 연구, 대전발전연구원
최예나. (2016). 사회적 자본이 지방정부 신뢰에 미치는 영향 연구: 주민들과 선출직 기관들간 소통의 조절효과를 중심으로.「지방정부연구」, 20(3), 69-88.
최유진. (2017). 도시어메니티의 지역경제 활성화 효과 분석: 우리나라 기초지방자치단체를 중심으로.「지방정부연구」, 20(4), 299-324.
하동현. (2017). 인구감소시대의 지역활성화와 지방분권-일본의 지방소멸론과 지방창생을 소재로.「한국지방행정학보 (KLAR)」, 14(3), 1-27.
하연섭. (2020). 한국 행정: 비교역사적 분석. 다산출판사
하혜수. (2017). 지방소멸시대의 지방자치 재검토-다양화와 차등화.「한국지방행정학보 (KLAR)」, 14(2), 1-24.
한국고용정보원. (2020). 지역고용동향브리프 2020년 여름호. 한국고용정보원
한상일·권소일. (2019). 사회적 기업 인지도와 사회적 자본의 사회적 기업 신뢰에 대한 효과 분석.「사회적경제와 정책연구」, 9(2), 33-55.

유한별(劉한별): 연세대학교에서 행정학 석사학위(석사, 2019, 연세대학교, 한국 4대강의 수질 환경 지표 변화 분석과 오염 위험 탐색에 관한 연구: 4대강 관련 수질 환경 정책의 영향을 중심으로)를 취득하고, 현재 연세대학교 행정학과 박사과정에 재학 중이다. 주요 관심분야는 공공관리, 규제정책, 갈등·환경 관리, 머신러닝 등이다(yhb5898@gmail.com).

탁근주(卓根柱): 연세대학교 응용통계학 석사학위(석사, 2021, 연세대학교, BERT 감성분석과 기술분석을 결합한 코스피지수 등락 예측 연구) 취득하였다. 주요 관심 분야는 머신러닝, 딥러닝, 자연어 처리 등이다(xkrrmswn@gmail.com).

문정승(文正承): 경기대학교 응용통계학 학사학위(학사, 2021, 경기대학교) 취득하였다. 주요 관심 분야는 머신러닝, 딥러닝, 빅데이터 분석 등이다(munmun2004@naver.com).

〈논문접수일: 2021. 1. 10 / 심사개시일: 2021. 1. 12 / 심사완료일: 2021. 2. 8〉

부록

기존 소멸지수로 도출된 소멸 고위험 지방자치단체명(총 62개)				
강원도 양양군	경상남도 합천군	경상북도 청송군	전라남도 장흥군	충청남도 부여군
강원도 영월군	경상북도 고령군	인천광역시 강화군	전라남도 진도군	충청남도 서천군
강원도 정선군	경상북도 군위군	인천광역시 옹진군	전라남도 함평군	충청남도 예산군
강원도 평창군	경상북도 문경시	전라남도 강진군	전라남도 해남군	충청남도 청양군
강원도 횡성군	경상북도 봉화군	전라남도 고흥군	전라북도 고창군	충청남도 태안군
경상남도 거창군	경상북도 상주시	전라남도 곡성군	전라북도 김제시	충청북도 괴산군
경상남도 고성군	경상북도 성주군	전라남도 구례군	전라북도 남원시	충청북도 단양군
경상남도 남해군	경상북도 영덕군	전라남도 담양군	전라북도 무주군	충청북도 보은군
경상남도 산청군	경상북도 영양군	전라남도 보성군	전라북도 부안군	충청북도 영동군
경상남도 의령군	경상북도 예천군	전라남도 신안군	전라북도 순창군	충청북도 옥천군
경상남도 창녕군	경상북도 울진군	전라남도 영광군	전라북도 임실군	
경상남도 하동군	경상북도 의성군	전라남도 완도군	전라북도 장수군	
경상남도 함양군	경상북도 청도군	전라남도 장성군	전라북도 진안군	

Abstract

A Study on the Factors and Overcoming Methods of Extinction of Provinces in Korea: The Exploration with Machine Learning methods

Yoo, Han Byeol
Tak, Keun Joo
Mun, Jeong Seung

This study aims to explore the extinction risk of local cities and counties in Korea. For analysis, it uses factors that affect the risk of extinction or the attractiveness of the region. and data is collected and organized in the KOSIS, and then, the improved extinction risk index is derived for local regions by reflecting population movement. the extinction risk index is utilized by dependent variables (target value). and we construct a machine learning analysis model with the independent variables(features) and dependent variables.

Machine learning models are GBM, RF, XGB, etc. and they predict and classify local decimation risk in a way that improves the performance of the model with enhancing methods like voting and ensemble. The results derive the local region with the highest risk of extinction and determine the factors that affect the local extinction risk.

As a result of the analysis, the prediction performance of the machine learning model built in this study was around 90% and 68 local regions were measured with the highest risk of extinction. Factors affecting these extinction risks were found to affect economic and industrial factors, and convenience factors such as cultural and medical facilities.

In conclusion, to overcome local extinction in the future, local governments in danger of extinction need to focus first on these economic and industrial factors improved, and if those factors are overcome, it is important to expand cultural and medical facilities to enhance local attractiveness.

Key Words: Local Extinction, Local Extinction Factors, Local Extinction Risk, Regional Attraction, Machine Learning