우 수 상

베이지안 가법 회귀 트리(BART) 및 심층 신경망을 기반으로 한 선제적 재활용률 분석 및 활용 방안

- 시차를 둔 영향요인 추출 및 예측 기반 정책 제안을 중심으로 -

| 김대영 |

본 논문 내용은 저자의 견해이며, 통계청 및 통계개발원 공식 견해와 일치하지 않을 수도 있습니다.

베이지안 가법 회귀 트리(BART) 및 심층 신경망을 기반으로 한 선제적 재활용률 분석 및 활용 방안

- 시차를 둔 영향요인 추출 및 예측 기반 정책 제안을 중심으로 -

김대영*

요 약

본 연구는 현 사회가 코로나-19 펜데믹을 거치며 직면하게 된 폐기물 차원상 탄소 중립 저해 현상에 집중하여, 이를 완화하기 위한 선제적 방안을 모색하는 것을 목적으로 설정하였다. 이에 환경부의 전국 폐기물 통계, 통계청 e-지방지표 통계 등을 기반으로 BART 모델과 DNN을 학습시킨 후, 각 모델을 전제로 특정 지역에서 발생할 재활용률을 약 1년 시차를 두고 예측하는 데 기여하는 주요 변수 및 변수별 주변 효과와 한계기여도를 추출, 이와 함께 재활용률 증진을 위한 선제적 개입 시 각 모델의 활용 방안을 제시하였다. 연구 결과 두 모형 모두에서 약 2%p의 오차 아래 유의한 재활용률 예측을 전개할 수 있음이 확인되었으며, 이때 공통적으로 자치구별 인당 공업지역 면적, 위탁급식영업체 수, 3인이상 세대 수가 향후 재활용률 증감에 유의한 영향을 미치는 것으로 나타났다. 본 연구는 선제적 재활용률 예측을 위한 실질적 모델과 함께 시차를 둔 영향요인을 추출한다는 점에서 의의를 갖는다.

주제어: 재활용률, 탄소중립, COVID-19, 포스트 코로나, DNN, BART, 폐기물 통계, MCMC, XAI, SHAP

* 연세대학교 상경대학 응용통계학과 학사과정

Ⅰ. 서 론

최근 코로나-19 펜데믹으로 인하여 다양한 영역에서 불가피한 일회용품 사용량이 증가하는 동시에 대부분의 거시적 사회정책이 경기 침체에 집중됨에 따라, 폐기물 차원에서의 탄소 중립 실현이 상당 수준 퇴보 및 저해되는 현상이 발생하였다. 구체적으로 코로나-19 펜데믹은 위생 및 방역 강화에 따른 의료용 일회용품 폐기물의 급증, 그리고 장기적인 차원의 사회적 거리 두기 강화로 인한 배달시장의 급격한 팽창을 발생시키며 일회용 포장 용기 등의 소비량을 폭증시킴에 따라 종합적인 차원에서 폐기물의 재활용 수준을 심각하게 저해하였다. 가령, 의료 환경에서는 매일 수십만 건수[1]의 COVID-19 유전자증폭방식(PCR) 검사 및 신속항원검사가 진행되는 동시에 격리시설 운영 및 중증환자 치료가 진행됨에 따라, 검체 채취 또는 치료 과정상 다양한 종류의 일회용 플라스틱 소재의 감염병(격리) 의료폐기물이 발생하여 왔으며, 일상적 차원에서는 방역 정책에 따라 폴리프로필렌(PP)을 주원료로 하는 다양한 종류의 일회용 마스크(KF-80, KF-94 등) 소비가 증폭됨에 따라 이전과 비교 시 재활용 수준상 상당 부분 악화가 발생하여 왔다. 한편, 코로나-19로 인한 사회적 거리 두기가 강화됨에 따라 배달시장에서는 음식 서비스 거래액 규모가 통계청 도소매·서비스 자료 기준 약 9조원(코로나 이전, 2019년: 9,735,362,000,000원)에서 약 26조원(코로나-19 펜데믹 시기, 2022년(추정치): 26,033,863,000,000원)으로 3배가량 성장하였으며, 이에 연쇄적으로 포장 용기, 특히 일회용 플라스틱 용기의 활용이 폭증하면서 폐기물의 재활용 차원상 악화가 급증한 것으로 파악된다.[2]

해당 현상은 단기적으로는 환경 파괴를 가속화하는 한편, 장기적으로는 사회의 지속가능성을 저해하는 요인으로서 작용할 수 있다는 점에서 정책적 차원의 선제적 개입이 시급한 영역이라 사료된다. 이때 현재까지의 폐기물 연구는 대부분의 경우 선제적 차원보다는 후속적 차원에서의 접근을, 분석모형의 설명력 및 예측력보다는 유의한 영향요인의 추출에 집중하는 접근을 강조해왔다는 점에서 현 사회가 직면하고 있는 탄소 중립 저해 현상을 해결하기 위한 대안을 제시하기에는 뚜렷한 한계가 존재한다고 판단된다. 본 연구는 이에 사후적 차원에서의 인과구조 파악 또는 패널

1) 질병관리청, COVID-19 DashBoard 기준. 2022년 2월 기준 월별 코로나-19 검사 현황: 총 10,312,319건. 2023년 2월 16일(논문 작성일) 기준 일일 코로나-19 검사 현황: 총 237,049건

2) 통계청, 『온라인쇼핑동향조사』, 취급상품범위별/상품군별거래액, 2023.02.01.

회귀분석 등을 전제로 한 재활용량 유의 영향요인 추출을 강조한 기존 선행연구와는 차별화되도록, 폐기물 처리 양상, 구체적으로는 지역구별 재활용률을 1년. 내지 2년 시차를 두고 예측할 수 있는 분석 모델을 구축하고 해당 모델을 기반으로 시차 전제하에 향후 재활용률에 영향을 미칠 수 있는 사회적 요소들을 분석하려는 시도를 전개할 필요성이 존재한다고 판단하였다. 본 연구는 구체적으로 머신러닝, 딥러닝, 그리고 베이지안 관점을 전제로, BART(Bayesian Additive Regression Tree) 모형과 심층신경망 모델을 활용한 자치구별 재활용률 예측 및 분석 모델을 구축, 이에 따른 주요 변수별 주변 효과(marginal effect)를 분석하는 동시에 교호작용을 고려한 변수별 한계 기여도를 XAI 기법을 통해 추출함으로써, 재활용률 악화 탐지 방안 및 정책적 개입이 필요할 것으로 예측되는 지역에 대한 접근 방안을 분석 및 제시하고자 하였다. 이때 분석의 범위는 서울특별시 25개 자치구로 제한하여 연구를 진행하였다.

Ⅱ. 이론적 배경

1. 재활용 및 폐기물 배출 영향 요인 관련 선행연구 검토

선행연구 검토 결과, 재활용 및 폐기물 배출과 관련된 다수의 연구는 대부분 크게 경제적 요인, 가구 형태, 그리고 거주지 특성을 중심으로 진행되어왔음을 확인하였다. 먼저 유광민·박정원(2022)는 LM 검정과 하우스만 검정을 기반으로 한 패널 분석을 통해 지역별 재정자주도, 외국인 인구, 그리고 도시면적 등 요인이 재활용품 배출량에 유의한 영향을 미침을 확인하였으며, 현승현·정지훈(2017)은 인구밀도, 아파트 거주세대 수, 그리고 인당 지방세 부담액이 재활용 처리량에 부의 영향을 미치는 유의한 요인임을 제시하였다. 다음으로 이소라(2018)는 구조방정식 모델을 기반으로 한 영향계수 분석을 통해 수요처의 확보 및 출고가격 안정성이 재활용 산업 활성화에 유의한 영향을 미침을 확인하였으며, 한준(2020)은 LMDI 방법론을 전제로 생활용품에서의 국제 유가 하락, 1인 가구 증가 등이 플라스틱 재활용 수준에 유의한 영향을 미치는 요인임을 제시하였다. 한편 김지욱·정의철(1998)은 순위 프로빗모형과 양측한계결절토빗모형을 기반으로 한 실증분석을 전개함으로써 가계 내 주부의 연령, 가계소득, 그리고 유보임금 등이 재활용 투입 시간에 유의한 영향을 미치는 요인임을 제시하였으며, 양진우·박해식(2003)은

경로분석 및 인과구조모형을 전제로 개인이 거주하는 주택형태, 혼인 여부, 그리고 개인의 심리·인지적 특성(재활용에 대한 부담감 및 분리배출 정보 인지 여부)이 재활용 수준에 유의한 영향을 미치는 변수로 작용함을 확인하였다. 마지막으로 홍성훈(2015)은 소득, 인구밀도, 그리고 종량제봉투의 가격이 재활용품 배출량에 유의한 영향을 미치는 요인임을 제시하였으며, 유두련(2002)은 주택형태(단독주택, 공동주택), 개인의 소득수준, 그리고 재활용 편의성이 재활용 행동 수준에 유의한 영향을 미침을 확인하였다.

선행연구 검토 결과, 재활용 및 폐기물 배출 영향 요인 관련 선행연구들은 인과구조모형 또는 회귀 모형 등을 전제로 개인의 소득 또는 경제수준, 연령, 거주형태, 가구원수, 원자재 가격, 인구 밀도, 지방세 부담 정도가 재활용률에 영향을 미치는 요인으로서 작용함을 확인하였다. 이를 통해 총 세 가지 부문에서 본 연구가 선행연구와의 차별성을 유지함을 확인하였다. 첫째, 본 연구는 반응변수인 재활용률과 설명변수 사이 일정 수준의 시차를 발생시켜 단순한 인과 구조가 아닌, 예측상 영향 변수를 추출하는 데 주력하고, 이후 영향별 방향성을 탐색한다는 점에서 설명력 및 예측력에 집중한다. 둘째, 본 연구는 전통적 통계 모형이 아닌 머신러닝 및 딥러닝 기법을 활용함으로 기존 접근과는 다른 관점의 재활용률 관련 지형 탐색을 전개한다는 점에서도 차별화된다. 셋째, 본 연구는 변수별 영향 분석에서 자주 간과되는 변수별 상호 작용을 면밀히 분석하고자 의사결정트리 기반의 모델을 활용하는 것은 물론, 변수별 주변 효과 및 변수별 상호 작용 전제 한계 기여도를 분석한다는 점에서도 또한 차별화된다. 이를 통해 본 연구에서는 기존에 시도하지 못했던 선제적 차원에서의 재활용률 악화 탐지 및 증진을 시도함으로써, 궁극적으로는 탄소 중립의 실현을 위한 깊이 있는 정책 제안을 진행하고자 한다.

2. XAI (eXplainable Artificial Intelligence) 분석

XAI란 eXplainable Artificial Intelligence의 약자로, 설명 가능한 인공지능을 통칭한다. XAI는 다양한 모델의 역량 및 학습 결과를 설명할 수 있다는 점에서 모델의 해석 가능성을 증진하는 수단으로서 활용되는데, 그 가운데 주된 의의는 블랙박스(black box)로 표현되는 딥러닝 모형 학습의 신뢰성을 높인다는 것에 있다(이재준 외 3인, 2021). 구체적으로, DNN(Deep Neural Network)이나 CNN(Convolutional Neural Network)과 같은 보편적 딥러닝 모델들은 기존에 존재하였던 통계 모형, 베이지안 네트워크, 또는 의사결정트리 모델들의 성능을

획기적으로 뛰어넘는 성능을 보여주지만, 어떻게 해당 모델에서 특정 예측 및 학습이 진행되었는지에 관하여는 기존 모형들에 비해 매우 불투명하다는 점을 고려할 때, 사용자의 측면에서는 해당 모델들을 쉽게 신뢰하기 어렵다는 특성을 갖는다. 또한, 딥러닝 모델들은 특정 특성 또는 설명변수가 반응변수의 값 또는 확률을 설명할 때 어느 방향(+, -)으로 어느 정도로 기여 하는지에 관한 뚜렷한 정보를 제공하지 못한다는 점에서도 상당 수준의 불투명성, 즉 한계를 갖는다. 이때 XAI는 변수별 중요도, 모델 학습 시 개별 표본별 특정 반응변수를 예측하는 데 각각의 변수가 기여하는 정도 등을 수치화 및 시각화함으로써 해당 단점들을 보완할 수 있게 된다(Gunning D et al. 2019, 안재현, 2020). 대표적인 XAI 기법으로는 섀플리 값을 전제로 한 SHAP, 국소적 신뢰도에 초점을 맞춘 LIME 등이 존재하며, 본 연구에서는 양적 반응변수를 전제로 분석을 진행한다는 점, 그리고 대부분의 설명 변수가 양적 변수라는 점, 두 가지 특성을 고려하여 여러 기법 가운데 SHAP을 기반으로 한 XAI 분석을 진행하였다.

Ⅲ. 연구 방법론

1. 분석 자료

본 연구는 환경부가 제공하는 전국 폐기물 발생 및 처리현황 통계(국가승인통계 제106029호), 통계청이 제공하는 KOSIS e-지방지표 통계, 그리고 서울특별시에서 제공하는 서울특별시 기본통계 등을 중심으로 진행되었다. 먼저 본 연구의 반응변수를 구성하는 '전국 폐기물 발생 및 처리현황 통계'란 폐기물관리법 제38조에 의거하여 환경부와 한국환경공단이 1년 단위로 행정구역별 폐기물 발생량 및 처리를 정리한 국가승인통계로, 특정 연도의 해당 폐기물 통계는 해당 연도의 익년 제3월부터 12월 사이에 조사된 다음, 익년도 말에 공표되는 방식으로 DB화 된다. 본 연구에서는 최대 2년 시차를 둔 설명변수를 활용함으로써 2023년 재활용률 등을 예측하고자, 현재 접근 가능한 설명변수별 최대 공시 연도(2020년~2021년)를 고려하여 해당 통계 가운데 2016년부터 2020년 데이터를 사용하는 방식으로 모델 학습을 위한 반응변수 데이터를 수집하였다. 다음으로, 본 연구의 설명변수를 구성하는 e-지방지표 통계와 서울특별시 기본통계 등은 앞선 선행연구에서 확인된 재활용량 영향요인(거주형태, 연령, 지방세 부담, 원자재 가격 등)에 직간접적으로

대응될 수 있는 요소 및 이전에 고려되지 않았던 지역적 특성(인당 도시지역·녹지지역 면적, 식품위생업 현황, 지역 내 재활용 가능 자원별 가격 등)에 대응될 수 있는, 차별화된 요소를 종합함으로써 요구되는 데이터를 수집하기 위해 활용되었다. 이를 통해 본 연구에서는 최종적으로 〈표 1〉과 같은 반응변수, 설명변수 데이터셋을 구축하였다. (이때 설명변수 집합 가운데 자치구별 인당 도시지역 면적, 인당 녹지면적, 자치구 내 일반음식점 수, 위탁급식업체 수, 즉석판매제조가공업체 수, 그리고 1인당 지방세 부담액 변수의 경우, 반응변수와 2년 시차를 발생시켰으며, 그 외 설명변수의 경우에는 1년 시차를 발생시켰다. 가령 재활용률

〈표 1〉 데이터셋을 구성한 변수 종류 및 요약

구분 (자치구별, 연도별)			변수 설명 및 (*학습 시 변수명)
반응변수	자치구별 총 재활용률		지역 내 총재활용량/총폐기물량(%)(*target)
설명변수	도시지역분포	인당 상업지역면적	지역 내 1인당 상업지역 면적(m^2) (*cm)
		인당 녹지지역면적	지역 내 1인당 녹지지역 면적(m^2) (*gre)
		인당 공업지역면적	지역 내 1인당 공업지역 면적(m^2) (*ind)
		인당 주거지역면적	지역 내 1인당 주거지역 면적(m^2) (*rd)
	인구구조 (연령)	유년부양비	0-14세 인구/생산가능인구 (*ysus)
		노년부양비	65세이상인구/0-14세 인구 (*esus)
	가구형태	1인 세대 수	1인 가구 규모 지표 (*oh)
		3인 이상 세대 수	3인 이상 가구 규모 지표 (*eh)
	식품위생업 현황	위탁급식영업체 수	지역 내 위탁급식업체 수 (*cif)
		즉석판매제조가공업체 수	지역 내 즉석판매제조가공업체 수 (*ipp)
		일반 음식점 수	지역 내 일반 음식점 수 (*rt)
	지방세 부담	인당 지방세부담액	자치구별 1인당 세금 부담 정도 (*ldp)
	재활용 가능 자원 원자재 수입 동향	펄프(LBKP)($/톤)	1톤별 LBKP 펄프 수입 가격 (*lbkp_r)
		펄프(NBKP)($/톤)	1톤별 NBKP 펄프 수입 가격 (*nbkp_r)
		두바이유($/bbl)	1배럴별 두바이유 수입 가격 (*dbo_r)
	재활용 품목별 가격 동향	PE재생FLAKE	폐플라스틱-pe재생flake 가격 (*peflake)
		PP재생FLAKE	폐플라스틱-pp재생flake 가격 (*ppflake)
		PE재생PELLET	폐플라스틱-pe재생pellet 가격 (*pepellet)
		PP재생PELLET	폐플라스틱-pp재생pellet 가격 (*pppellet)
		폐금속캔(철캔)	폐금속캔(철캔) 가격 (*sc)
	구매력·경제 수준	소비자물가 등락률	서울시 CPI 등락률 (*cpi_p)

*환경부 통계, 통계청 KOSIS e지방지표, 서울열린데이터광장 서울특별시 기본통계

데이터가 2020년 데이터일 경우, 자치구 내 1인 가구 수는 2019년 데이터. 즉석판매제조가공업체 수는 2018년 데이터.)
(총 5개년도 22개 설명변수 × 25개 자치구별 거주지 변수: 125행 x 22열).

2. 주요 변수

1) 반응변수

본 연구에서의 반응변수로는 특정 연도의 자치구별 총 재활용률(=총 재활용량/총 폐기물 배출량)이 활용되었다. 이때 반응변수 도출에는 앞선 목적을 고려하여 2016년~2020년 폐기물 처리 데이터(국가승인통계 106029호)가 활용되었으며, 본 연구의 분석 범위는 서울특별시로 한정되었다는 점에서 총 125개(5개년도×25개 자치구)의 반응변수 데이터셋이 모델 학습에 활용되었다.

2) 설명변수

본 연구에서는 설명변수로서 크게 도시지역 분포 변수, 인구구조(연령) 변수, 가구형태 변수, 식품위생업 현황 변수, 지방세 부담 변수, 재활용 가능 자원 원자재별 수입 동향 변수, 재활용 품목별 가격 동향 변수, 그리고 구매력·경제 수준 변수를 설정하였다. 이때 도시지역 분포 변수와 식품위생업 현황, 그리고 재활용 품목별 가격 동향은 지역별 산업 지형 및 재활용품 시장 지형을 반영하는 요인에 대응되는 특성이자 일회용품 발생 가능성을 높이는 물리적·경제적 잠재 요인에 대응되는 특성으로서 선정되었으며, 나머지 설명변수의 경우에는 순차적으로 각각 선행연구에서 확인된 영향요인인 연령, 거주형태, 지방세 부담, 원자재 가격, 그리고 실질적 경제수준에 직간접적으로 대응될 수 있는 특성으로서 선정되었다.

또한, 각 설명변수는 다음과 같은 하위 범주로 재분류되었다. 먼저 도시지역 분포 변수는 인당 상업지역, 공업지역, 주거지역, 그리고 녹지지역 면적으로 소분류되었으며, 가구형태 변수는 1인 세대 수와 3인 이상 세대 수로 소분류 되었다. 다음으로, 식품위생업 현황 변수는 위탁급식영업체 수, 즉석판매제조가공업체 수, 그리고 일반 음식점 수로 소분류되었으며, 재활용 가능 자원에 대한 원자재별 수입 동향 변수는 펄프(LBKP, NBKP) 톤당 가격과 배럴 당 두바이유 가격으로, 재활용 품목별 가격 동향 변수는 PP, PEflake 및 pellet과 폐금속캔(철캔)으로 소분류 되었다.

3. 분석 방법

1) MCMC 및 상호정보량(MI)의 변화점을 활용한 변수선택(차원축소)

본 연구에서는 모델의 학습을 위해 훈련 데이터 80%, 평가 데이터 20%의 비율로 앞서 구축한 데이터셋에서 홀드아웃 샘플링(Holdout Sampling)을 시행하였으며, 이때 샘플링 수단으로는 python sklearn.model_selection 패키지의 train_test_split()함수를 활용하였다(샘플링을 진행한 후에는 모델의 학습 효율성을 증진하기 위하여 각 설명변수에 대한 표준화 진행). 한편, 본 연구에서는 모델의 학습 및 평가에 앞서. 주어진 설명변수 집합 내 차원 축소를 실현하기 위해, 상호정보량을 기준으로 설명변수들을 중요도를 분석함으로써 변수 선택을 시행하였다. 상호정보량(Mutual Information)이란, Shannon(1948)이 제시한 정보엔트로피를 기반으로 하여 변수 간 상관성을 정량화한 값으로, 특정 변수의 데이터에 따라 다른 변수의 불확실성이 감소하는 수준을 측정하여 설명변수의 중요도를 제시한다는 점에서 주요 의의를 갖는다 (수식 1). 본 연구에서는 상호정보량을 내림차순으로 정리한 후, 더 이상 상호정보량이 유의미하게 감소하지 않는 변화점을 MCMC를 통해 추출하여 해당 지점까지의 변수들만을 유효 변수로서 선정하는 방식으로 차원 축소를 시행하였다.

$$
\begin{aligned}
&Information\ Entropy:\ H(X) = -\sum_{x \in X} p(x)\log p(x) \\
&Mutual\ Information:\ \sum_{y \in Y}\sum_{x \in X} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \\
&\qquad = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) \qquad \cdots\cdots\cdots(1) \\
&p(x): marginal\ pmf\ (pdf)\ of\ x
\end{aligned}
$$

(자료 : 추한경 외 4인, 2018)

2) 연구 모형(모델)의 구축·학습·평가

(1) BART 모델의 정의, 구축 및 활용

본 연구에서는 첫 번째 분석 모형으로서 먼저 베이지안 가법 회귀 트리(BART) 모형을 구축 및 학습하였다. BART(Bayesian Additive Regression Tree)란, 비모수적 베이지안 회귀 방법론 가운데 하나로, 머신러닝 기법인 의사결정 트리(decision tree)의 집합을 베이지안 backfitting MCMC 샘플링과 연결함으로써.

반응변수를 정확하게 예측하기 위한 최적 회귀 트리 조합에 대한 사후분포 표본을 반환받는 방식의 모델을 일컫는다. 이때 회귀 트리 조합을 구성하는 개별 회귀 트리는 반응변수 전체가 아닌 일부를 각각 설명하는 부분적 접근(sum-of-trees)을 전개함으로써, 전체 반응변수에 대한 개별 트리 적합을 전제로 사후적 평균치를 구하는 접근법과는 개념적으로 구별된다(수식 2)(Chipman et al., 2010).

$$BART\ model : Sum-of-Trees$$

$Y = BART(X) + \varepsilon$ $(\mu = BART(X), \varepsilon \sim N(0,\sigma^2))$이라 할 때,

$$Y = f(X) + \varepsilon \qquad X = < x_1, x_2, \cdots x_p >, \ \varepsilon \sim N(0,\sigma^2) \qquad \cdots\cdots\cdots(2)$$

$E(Y \mid X) = f(X) \approx h(X) = \sum_{j=1}^{m} g(X;T_j,M_j) = \sum_{j=1}^{m} \mu_{ij}$ (g_j : Y에 대해 j번째 트리가 설명하는 부분)

(m : 회귀트리 개수, T_j : j번째 트리, M_j : j번째 트리의 최종마디에 대응되는 μ_{ij})

구체적으로 BART 모델은 (수식 2)의 Sum-of-Trees 접근 상 발생하는 모든 파라미터별로 사전분포(prior for $\{(T_1,M_1),(T_2,M_2),\cdots,(T_m,M_m),\sigma\}$)의 존재를 가정한 다음, 총 네 가지 사전 규제(regularization prior) 원칙을 전제로 backfitting MCMC 샘플링을 진행하여 유의한 사후 표본을 확보한다(수식 2). 첫째, BART는 Sum-of-Trees 모델상 개별 파라미터 사이의 독립성(independence)과 대칭성(symmetry)을 가정한다. 둘째, BART는 개별 트리의 사전분포($p(T_j)$) 가정 시, 어떠한 노드가 깊이 d에서 해당 트리의 최종 마디(terminal node)로 분기되지 않을 확률을 $\alpha(1+d)^{\beta}$ $(\alpha \in (0,1),\ \beta \in [0,\infty))$이라 정의한 후, 회귀 트리의 각 interior 노드에서의 특성 부여 및 분기 규칙 부여에 대한 분포로서 균등분포 등을 가정함으로써 개별 트리의 구조를 최대한 소규모로 생성하는 것을 시도한다. 셋째, BART는 $\mu_{ij} \mid T_j$에 대한 조건부 사전분포에 대하여 공액사전분포(conjugate prior distribution)인 $Normal(\mu_\mu,\sigma_\mu^2)$ 등을 활용하며, 트리 적합 상의 과밀화 또는 과분산을 방지하기 위해 반응변수 Y에 대해 최솟값이 -0.5, 최댓값이 0.5가 되도록 scaling을 실시한 후, 해당 자료를 전제로 초모수인 μ_μ와σ_μ^2 값이 $m\mu_\mu - k\sqrt{m}\sigma_\mu = y_{\min}$, $m\mu_\mu + k\sqrt{m}\sigma_\mu = y_{\max}$, 두 식을 만족하도록 설정함으로써 최종적으로는 반응변수 Y의 수치 가운데 j번째 트리인 T_j에 의해 설명되는 부분인 $\mu_{ij} \equiv g(X;T_j,M_j)$가 $N(0,\sigma_\mu^2)$을 따르도록 규제한다. 마지막으로 넷째, BART는 σ에 대한 사전분포로서는

공액사전분포인 scaled 역-카이제곱분포를 설정한 다음, 초모수인 v와 s가 각각 $3 \le v \le 10$, $P(\sigma < \hat{\sigma}) = q\,(q = .70, .90, .99$ 등$)$[3]를 만족하도록 선정하여 $P(\sigma)$의 구체적 형태를 확정한다. 해당 규제들을 통해 BART 모형은 일부 트리가 반응변수 Y의 대부분을 설명하고 나머지 트리는 국소적인 부분만을 탐색하는 식의 모형 적합 가능성을 제거함으로써 모델 자체의 대표성 및 유연성을 강화한다.

Bayesian Approach for Parameters of BART

$$P((T_1,M_1),(T_2,M_2),\cdots,(T_m,M_m),\sigma \mid Y) \propto P(Y \mid (T_1,M_1)\cdots(T_m,M_m),\sigma)P(T_1,\cdots\sigma)$$

$$(\because P(\theta \mid y) = \frac{P(y \mid \theta)P(\theta)}{\int P(y \mid \theta)P(\theta)\,d\theta}(P(y \mid \theta): Likelihood, P(\theta): prior\,distribution))$$

$$P(T_1,\cdots\sigma) = P(\sigma)\prod_{j=1}^{m}P(T_j,M_j)\,(\because Independence) = P(\sigma)\prod_{j=1}^{m}P(M_j \mid T_j)P(T_j)$$

$$= P(\sigma)\prod_{j=1}^{m}P(T_j)\prod_{i=1}^{}P(\mu_{ij} \mid T_j)\,(\because symmetry) \qquad \cdots\cdots\cdots(3)$$

$$Prior\,Distributions\,by\,regulariation\,prior: P(T_j), P(\mu_{ij}|\,T_j) \sim N(\mu_\mu,\sigma_\mu^2), P(\sigma) \sim Inv-\chi^2(v,\lambda)$$

다음으로, BART 모델은 해당 사전 규제 가정들을 전제로 backfitting MCMC를 시행하며, 이때 Gibbs 샘플러의 활용을 가정할 경우, 해당 MCMC 샘플링은 $p((T_j,M_j) \mid T_{(j)},M_{(j)},\sigma) \equiv P((T_j,M_j) \mid R_j,\sigma)\,(R_j \equiv Y - \sum_{k \neq j} g(X;T_k,M_k))$에서의 연속적 표본 추출을 필요로 하며, 해당 분포는 각각 Metropolis-Hastings 알고리즘을 전제로 한 $p(T_j \mid R_j,\sigma)$에서의 표본 생성과 사전 규제에 따라 생성되는 $p(M_j \mid T_j,R_j,\sigma)$에서의 순차적 표본 생성으로 분해된다. BART에서는 해당 MCMC 샘플링을 정해진 횟수만큼(=K번) 시행하여, 최종적으로는 수렴 전제하에 m개 트리 및 σ에 대한 총 K개의 사후 표본 $((T_1^{(1)},M_1^{(1)}),\cdots(T_m^{(1)},M_m^{(1)}),\sigma))$, $\cdots((T_1^{(K)},M_1^{(K)}),\cdots(T_m^{(K)},M_m^{(K)}),\sigma))$을 반환받는다. 이에 따라 특정 데이터(*)에 대한 반응변수 Y의 값에 대한 적합 또는 예측은 $E(f(x) \mid y)$로 대표되며, 해당 수치는 앞서 생성된 K개의 사후분포 표본 집합에 따라 계산되는 $f_1^*, f_2^*, \cdots, f_k^*$ $(f^*(X^*) = \sum_{j=1}^{m} g(X^*;T_j,M_j))$의 평균치 또는 중앙값에 대응됨으로써 최종적으로

3) 여기서 $\hat{\sigma}$는 naive specification(Y의 표본분산) 또는 linear model specification(X와 Y에 대한 OLS 회귀 전제 시 잔차의 표본분산)을 전제로 도출된다.

정리된다. 이때, BART는 $f_1^*, f_2^*, \cdots, f_k^*$를 활용하여 빈도수를 전제로 변수별 부분 의존도 함수를 계산하며, i번째 변수가 트리의 분기 규칙을 형성하는 데 사용된 빈도수에 따른 비율을 z_{ik}라 정의한 다음, 해당 수치의 단순 평균 또는 초기 노드 분기에 가중치를 둔 가중평균을 활용함으로써 특정 변수의 중요도를 수치적으로 정리하여 모델의 해석 가능성을 높인다(Chipman et al., 2010).

본 연구에서는 프로그래밍 언어 Python의 pymc 및 pymc_bart 패키지를 활용하여 BART 모델을 구축 및 학습하였으며, 이때 σ를 위한 사전분포로는 앞선 사전 규제 조건과 역-카이제곱 분포의 성질을 활용하여, 3부터 10까지의 값 가운데 중간치인 6을 v로 가정한 다음 이에 대응되도록 Inverse-Gamma (α=3, β=16) 분포를 가정한 후, 두 개의 병렬 MCMC 체인에서 각각 총 3000번의 표본 생성, 1000번의 튜닝을 진행함으로써 사후분포 표본을 추출하였다. 이때 각 표본과 관련한 사슬별 수렴 진단은 대응되는 r_hat($\hat{R}$) 값을 전제로 진행되었다. 이후 본 연구에서는 형성된 6000개 표본을 전제로 ppc plot 생성, 부분 의존성 plot 생성, 변수 중요도 그래프 생성, 그리고 평가 데이터에 대한 예측치 생성을 진행하였으며, 이때 모델의 설명력 및 예측력에 대한 판단 기준으로는 평가 데이터에 따른 예측치와 실제 데이터 사이 RMSE와 MAE를 활용하였다.

(2) 심층 신경망(DNN) 모델의 구축 및 활용

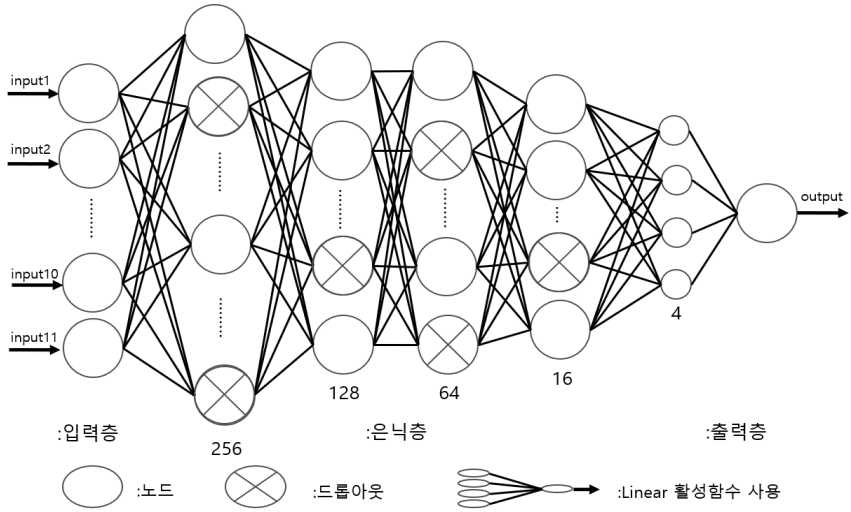
다음으로, 본 연구에서는 두 번째 분석모형으로서 딥러닝을 기반으로 한 심층 신경망(DNN)을 활용하였다(그림 1 참조). 이때 신경망(DNN)의 구조로는 Python의 Tensorflow와 Keras 라이브러리를 활용하여 총 5개의 은닉층을 설정하였으며(순차적으로 각각 2의 k 승(k= 자연수)인 256, 128, 64, 16, 4개 노드 생성), 각 은닉층과 관련하여서는 과적합을 방지하기 위해 0.5의 드롭아웃을 설정한 다음, 배치 정규화 및 L2 규제를 함께 적용하였다. 다음으로, 본 연구의 DNN을 위한 활성화 함수로는 swish 함수를 활용하여 모델 구조를 구축하였으며[4], 경사 소멸

4) SWISH(Self-gated activation function) 활성화 함수는 Google Brain의 연구진이 신경망 학습 시 발생하는 기울기 소멸 문제를 해결하기 위해 relu 및 sigmoid 활성함수의 특성을 종합함으로써 구축한 활성함수로, 모델 성능을 높이는 동시에 함수가 위로 유계되어 있지 않아 학습 과정에서의 포화 효과를 제한한다는 특성을 지니고 있다는 측면에서 효율적인 신경망 구축 실현하는 데에 기여할 수 있는 수단으로 높이 평가된다(Faiama A.;Pethe A., 2022, Ramachandran P., Zoph B., Le QV., 2017).

문제를 방지하는 동시에 해당 모델을 최적화하고자 분산 조정 기반 가중치 초기화 수단 가운데 하나인 He 초기화를 활용하여 모델 성능을 개선하는 것을 시도하였다 (수식 4) (Narkhede, M.V. et al., 2022, He et al. 2015). 한편 DNN 학습 방식으로는 손실함수로서 평균제곱오차(MSE)를, 평가도구로는 평균절대값오차 (MAE)를 설정하여 학습을 전개하였으며, 에포크(epoch)로는 5000번을 설정하여 학습을 진행하였다. 이후 본 연구에서는 최적화 도구(optimizer)와 배치 크기 (batch size)를 하이퍼파라미터로 설정한 다음, 최적화 도구 집합으로는 {adam, rmsprop}을, 배치 크기 집합으로는 {10, 20, 30, 40, 50}을 가정한 후 MAE를 기준으로 하여 해당 지형 내 최적 파라미터 조합을 선택하는 방식으로 모델 학습을 전개하였다. 마지막으로 본 연구에서는 앞서 생성한 평가 데이터를 전제로 최종 모델에 대한 평가를 진행하였으며, 평가도구로는 학습 시와 마찬가지로 평균 절댓값 오차(MAE)를 활용하였다.

$$He\ Initalization\ Using\ Normal\ Distribution\ (random\ sampling).$$
$$Weight \sim Normal(0, \frac{2}{f_{in}}).\ standard\ deviation(Weight) = \sqrt{\frac{2}{f_{in}}} \quad \cdots\cdots\cdots(4)$$
$$f_{in} : Previous\ node\ inputs.$$

〈그림 1〉 본 연구의 심층 신경망(DNN) 기본 구조 예시(input은 변수선택 통해 추출된 변수 11개, 은닉층은 5개(은닉층별 노드 수: 256, 128, 64, 16, 4))



3) 모델의 해석 및 대응 전략 분석

(1) BART 모델 해석: PPC plot, 특정 중요도 그래프, 그리고 부분 의존성 그래프(PDP)를 활용한 모델 해석 및 반응변수에 대한 주요 설명변수별 주변 효과 파악.

본 연구에서는 앞서 가정된 두 모형을 학습한 다음, RMSE, MAE 등의 지표를 기반으로 각 모델이 충분한 설명력 및 예측력을 갖추었음을 전제로 하여 재활용률에 대한 데이터 분석을 재개하였다. 이때 적합된 BART 모형의 해석과 관련하여서는 크게 Posterior Predictive Check 그래프, 특성 중요도 그래프, 그리고 부분 의존성 그래프(Partial Dependece Plot)가 활용되었다. 구체적으로 본 연구에서는 특성 중요도 그래프에서 확인되는 엘보우 포인트를 확인, 해당 포인트까지의 변수를 재활용률 예측에 영향을 미치는 주요 변수로 별도 분류한 후, 해당 변수들에 대한 PDP 그래프를 시각화하였다. 이때 본 연구에서는 부분 의존성 그래프에서 확인되는 주변 효과의 방향성을 전제로 주요 변수와 재활용률 증감 사이 연관을 구체화하는 동시에, 특별한 주변 효과가 발견되지 않는 주요 변수와 관련하여서는 이후 진행되는 SHAP 분석을 통해 해당 변수와 주요 설명변수 집합 내 기타 변수와의 연관을 파악하여 각 변수의 한계 기여도 및 타 변수와의 상호작용을 추출하였다.

(2) SHAP 기반 DNN 결과 해석 및 상호 작용을 고려한 변수별 한계 기여도 파악

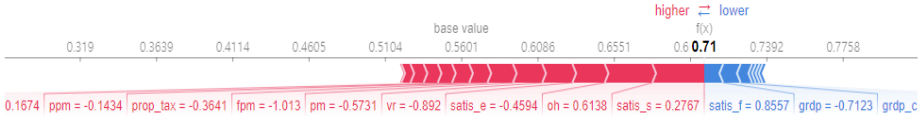
한편 학습된 심층신경망과 관련하여서는 (MAE 등 기준 하) 높은 수준의 정확도를 확보했다는 전제하에 해당 모델에 대한 XAI 분석을 진행함으로써 상호작용 전제하 설명변수별 영향을 분석하였다. XAI 분석으로는 섀플리(Shapley) 값을 사용하기 위한 수단으로서 python의 SHAP 라이브러리를 활용하였으며. 본 연구에서는 구체적으로 KernelExplainer() 함수를 통해 각 변수의 중요도 파악, 그리고 각 설명변수가 연속형 반응변수 값에 기여하는 방식, 즉, 각 변수가 재활용률의 증진에 영향을 미치는 방향성(+,-)에 대한 파악을 진행하였다. 여기서 SHAP이란 Lundberg S.M. et al(2017)이 최초로 제시한 개념으로, Shapley(1950)가 제안한 게임이론에서의 shapley 값, 즉 게임의 개별 참가자별 한계 기여도 개념을 머신러닝 및 딥러닝 모델 분석에 적용함으로써. 반응변수의 값에 개별 특성(설명변수)이 기여하는 정도를 분석할 수 있도록 한 XAI 기법을 말한다. 이때 개별 특성의

기여도, 즉 shapley 값은 특성 집합을 연합 N, 연합의 부분집합을 S, 분석 대상이 되는 특성을 i라고 정의할 때, (N의 전체 원소에 대하여 생성한 순열별 한계 기여도 계산을 전제로) 다음과 같은 수식으로서 정리되며(수식 5), 이때의 shapley 값은 학습이 완료된 모델을 기반으로 각 샘플별 모델 예측값(본 연구에서는 자치구별 재활용률)에 대한 개별 변수의 기여도 정도를 변수 간 상호 작용 전제하에 수치적인 자료로 제시한다는 점에서 모델에 대한 불투명성을 크게 줄인다는 의의가 존재한다 (예시: 그림 2).

$$Shapley\ \ Value\ \text{for}\ feature\ \{i\} \equiv marginal\ contribution\ of\ feature\ \{i\}$$
$$\phi_i(N,v) = \frac{1}{N!}\sum_{S \subseteq N \setminus \{i\}} |S|!(|N| - |S| - 1)!(v(S \cup \{i\}) - v(S))$$
$$(*\ v : MODEL,\ N \setminus \{i\} : feature\ \{i\}\ removed\ from\ set\ N). \qquad \cdots\cdots\cdots(5)$$

(자료 : *Rothman D.*, 2022)

〈그림 2〉 (예시) 이항 반응변수 기반 DNN 모델 학습 후 특정 샘플에 대한 SHAP 분석



*해석 예시: oh는 0.6138의 shapley 값을, grdp는 -0.7123의 shapley 값을 가짐으로써, 해당 샘플상 반응변수 1이 발생할 확률 예측치(0.71) 도출에 각각 긍정적, 부정적 역할을 수행하였음을 확인 가능.

이를 통해 본 연구에서는 앞서 BART에서 확인된 변수 중요도와 SHAP 분석 결과를 비교 및 종합함으로써, 구축된 설명변수 집합 가운데 강하게 대두되는 주요 변수 및 해당 변수별 주변 효과와 한계 기여도를 함께 고려하여 재활용률 예측상 유용한 신호를 제공할 수 있는 변수별 움직임을 입체적으로 분석하였다.

(3) 대응 방안 분석

본 연구에서는 BART 모델 적합 결과와 DNN을 활용한 XAI 분석을 종합하여 특정 자치구에서의 향후(약 1년에서 2년 시차) 재활용률에 영향을 미친다고 파악되는 주요 변수들을 추출하고, 해당 주요 변수 간 상관관계를 활용하여 최종적으로는

특정 자치구 내에서의 재활용률 변화 탐지방안, 그리고 재활용률 증진을 위한 정책 차원의 대응 진행 시 본 연구에서 구축한 두 모델의 구체적 활용 방안을 제시하고자 하였다. 이때 본 연구는 탐지방안 및 정책적 차원의 대응에 대한 구체성을 더하고자 일반적인 차원에서의 선제적 재활용률 악화 대응 방안을 제시함과 함께, 활용 방안 상의 선례를 제시하는 차원에서 앞서 구축된 두 모델을 전제로 서울시 25개 자치구별 2023년 재활용률을 예측하여 정책적 개입이 필요한 위험 자치구를 추출, 해당 자치구들에 대한 맞춤형 재활용 증진 정책 제안을 진행하여 요약하였다.

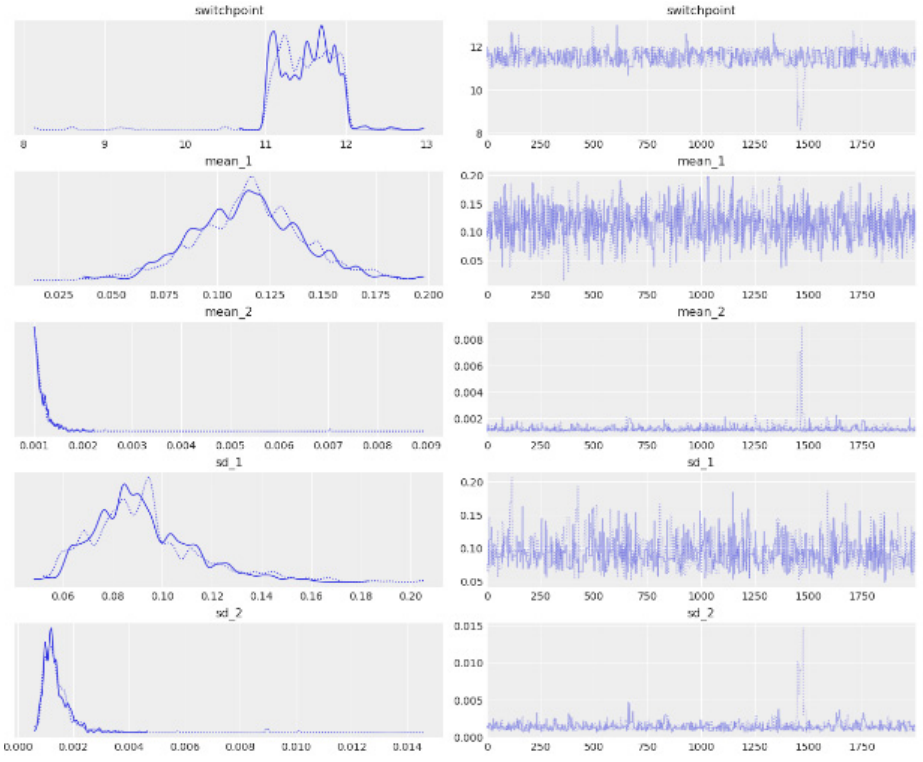
Ⅳ. 연구 결과

1. 상호정보량(MI) 변화점 기반 변수 선택에 따른 차원 축소 결과

먼저 설명변수 집합과 관련하여 각 변수별 상호정보량을 계산하여 내림차순으로 정리한 다음, 해당 데이터를 전제로 주요 변수 선택을 진행하였다. 이때, 본 연구에서는 해당 내림차순 데이터 열에서 상호정보량의 양상이 변화하는 지점이 존재할 것으로 가정하여 MCMC 샘플링을 진행하였으며, 정규분포 전제하에 변화점 τ의 사전분포로서 균등분포(≡Uniform(0, 21))를, 변화점 전후 두 집단별 $\mu_{1(2)}$의 사전분포로서는 균등분포(≡Uniform(0.001, 0.2))를, $\sigma_{1(2)}$의 사전분포로는 지수분포(≡Exponential(λ=1/0.08)를 설정한 후 MH-알고리즘을 사슬별로 2000번 반복한 결과, (그림 3)과 같은 결론이 도출되었다. 먼저 $\tau, \mu_1, \mu_2, sd_1, sd_2$의 사후분포 평균으로는 소수점 셋째 자리까지 반올림 시 각각 11.496, 0.115, 0.001, 0.092, 0.001이 도출되었으며, 표준편차로는 각각 0.385, 0.026, 0.000, 0.021, 0.001이 도출되었다. 다음으로, $\tau, \mu_1, \mu_2, sd_1, sd_2$에 대한 사슬별 $\hat{R}$ 수치로는 각각 1.002, 1.001, 1.012, 1.004, 1.008이 도출되었으며, 이에 모든 모수별 사슬상 MCMC 수렴이 확인됨에 따라 해당 분석의 유의성이 보장됨을 확인하였다. 따라서 본 연구에서는 전체 설명변수 가운데 11개까지의 변수가 재활용률 예측에 주요한 영향을 미치는 변수라 판단하였으며, 이때 해당 변수로는 각각 자치구별 1인 세대 수, 인당 공업지역 면적, 3인 이상 세대 수, 인당 녹지면적, 위탁 급식

영업체 수, 일반음식점 수, 인당 상업지역 면적, 유년부양비, 즉석판매제조가공업, 인당 지방세 부담액, 그리고 인당 주거지역 면적 변수가 존재함이 확인되었다.

〈그림 3〉 MH-알고리즘을 전제로 한 상호정보량 변화점 MCMC 샘플링

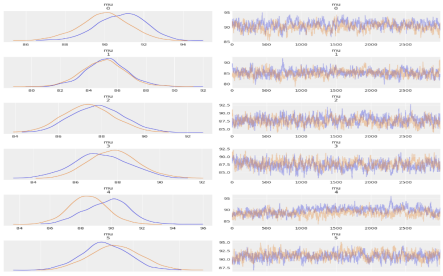


2. BART(Bayesian Additive Regression Tree) 학습 및 평가, 해석 결과

다음으로, 총 30개의 회귀 트리를 전제로 학습 데이터셋에 대하여 BART를 학습한 결과, (그림 4)와 같이 MCMC 사후분포 표본이 생성되었으며, 사슬별 $\hat{R}$ 값을 정리한 결과, 4개 훈련 데이터를 제외한 모든 경우 1.2 미만의 수치(수렴)가 발생하였으며, 이에 따라 본 연구에서 활용된 30개의 회귀 트리는 주어진 데이터셋에 대하여 높은 적합도를 보인다고 판단되었다. 또한, 훈련 데이터셋에 대한 PPC plot을 시각화한 결과, (그림 5)와 같은 결과가 도출되었으며, 이때 6000개의 파라미터 표본 집합

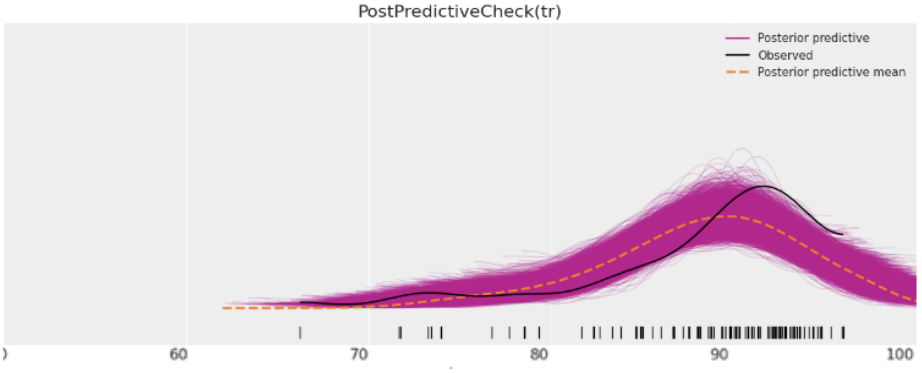
$((T_1^{(1)}, M_1^{(1)}), \cdots, (T_{30}^{(1)}, M_{30}^{(1)}), \sigma^{(1)}))$, $\cdots$, $((T_1^{(6000)}, \cdots, (T_{30}^{(6000)}, M_{30}^{(6000)}), \sigma^{(6000)}))$에 따른 반응변수별 개별 예측치를 계산한 다음, 해당 예측치들을 전제로 한 확률밀도함수 6000개를 중첩했을 때 발생하는 일반적 양상이 실제 반응변수 데이터의 밀도함수와 특이점 없이 매우 유사한 것으로 판단됨에 따라, 본 MCMC 샘플링을 통해 학습된 BART 모델은 주어진 훈련 데이터에 대한 높은 적합 및 설명력을 내포하고 있음을 다시 한번 확인하였다. 이때 훈련 데이터에 대한 RMSE와 MAE로는 (사후표본에 따른 반응변수 예측치 산술평균을 기준으로 계산 시) 각각 3.257%p, 그리고 2.365%p가 도출되었다. 다음으로, 도출된 6000개 사후 표본에 따른 BART 모델을 앞서 구축한 평가 데이터를 활용하여 평가한 결과, RMSE로는 3.464%p, MAE로는 2.666%p가 도출됨에 따라 본 BART 모델은 예측력의 차원에서도 우수한 성능을 보이는 것이 확인되었으며, 이때 사후 표본의 중첩을 기반으로 한 확률밀도함수를 시각화할 시 (그림 6)과 같은 양상을 보임이 확인되었다.

〈그림 4〉 좌: BART MCMC 샘플링 시각화 일부, 우: 훈련 데이터셋에 대한 체인별 요약

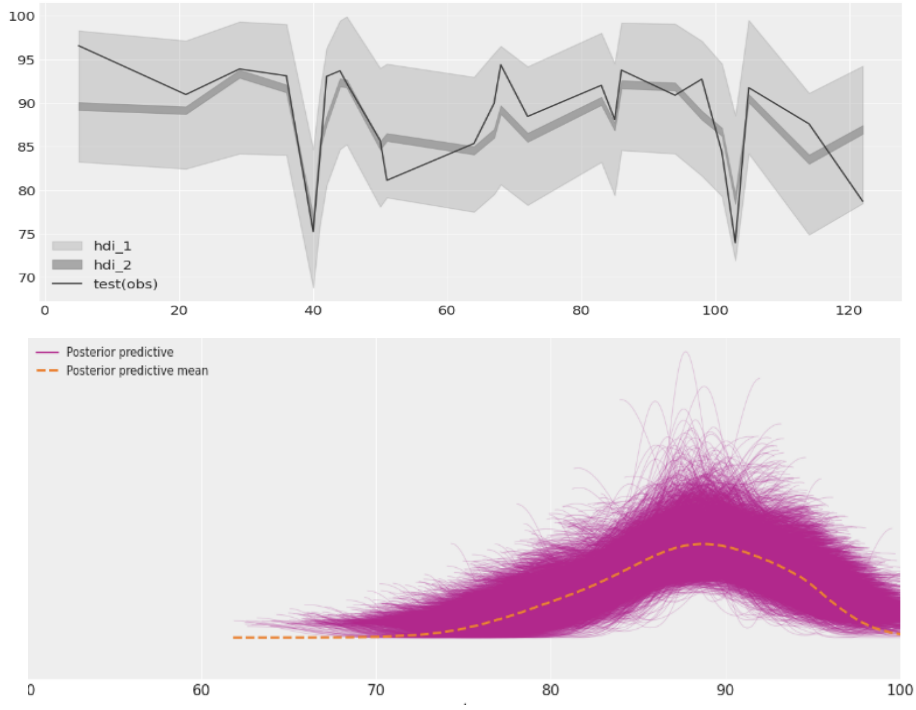


	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
mu[0]	90.428	1.321	87.809	92.807	0.268	0.192	25.036	362.330	1.087
mu[1]	85.144	1.504	82.324	88.062	0.091	0.064	274.690	568.628	1.017
mu[2]	87.620	1.202	85.392	89.841	0.071	0.050	285.571	703.212	1.026
mu[3]	87.510	1.301	84.970	89.885	0.130	0.092	98.439	519.336	1.040
mu[4]	89.098	1.500	86.385	91.837	0.551	0.407	7.606	82.744	1.194
...	...	...	...	...	...	...	...	...	...
mu[96]	89.552	1.240	87.212	91.952	0.073	0.051	291.279	611.346	1.014
mu[97]	92.344	1.111	90.278	94.476	0.101	0.072	121.765	415.206	1.032
mu[98]	83.148	1.564	80.228	86.051	0.101	0.072	237.761	575.817	1.006
mu[99]	89.792	1.169	87.469	91.922	0.074	0.053	246.841	521.031	1.020
sd_	3.582	0.404	2.847	4.334	0.130	0.095	9.501	155.789	1.155

〈그림 5〉 Posterior Predictive Check Plot for Train Data



〈그림 6〉 상: 평가 데이터에 대한 BART 모형 예측 HDI(10% 구간, 95% 구간)
하: 평가 데이터에 대한 BART 모형 예측에 따른 PPC plot

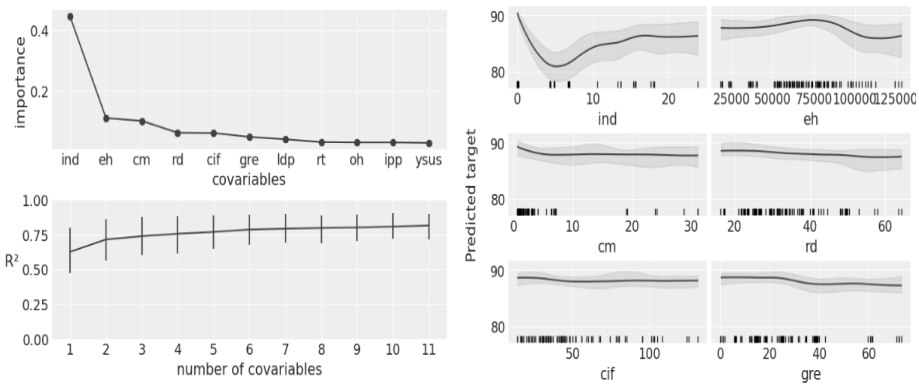


*HDI 그래프(상): hdi_1은 95% HDI, hdi_2는 10% HDI, X축은 평가 데이터셋별 인덱스(index). 이때 본 연구는 2016년부터 2020년까지의 연속적 자료를 시계열 자료를 사용한다는 점에서, 위 그래프 상 결과는 BART 모델이 단순히 높은 예측 성능을 보이는 것을 넘어 동일 자치구의 서로 다른 시점 고려 시 시계열 추세상 변화도 잘 적합하였음을 제시한다.

다음으로, 본 연구에서는 BART 모델에서 확인된 설명력 및 예측 성능을 전제로 개별 변수에 대한 중요도 및 주변 효과를 분석하였다(그림7). 이때 변수 중요도 그래프를 시각화한 결과, 3인 이상 세대 수 변수에서 엘보우 포인트가 발생하는 현상이 확인되었으며, 이때 해당 지점까지의 변수를 주요 변수로 선정하기에는 변수 개수가 매우 과소하다는 판단하에 나머지 변수 가운데에서 상위 50%를 추가로 선정하는 방식으로 주요 변수 선정을 선정함으로써 ind(1인당 공업면적), eh(3인 이상 세대 수), cm(1인당 상업면적), rd(1인당 주거면적), cif(위탁급식 영업체 수), 그리고 gre(1인당 녹지면적), 총 여섯 가지 변수를 전제로 주변 효과

(marginal effect) 분석을 전개하였다. 먼저 1인당 공업지역 면적 변수의 경우 인당 약 $5m^2$까지는 해당 수치가 증가할수록 향후 해당 지역의 재활용률 수치가 감소하는 현상이 발생하나, 그 이후부터는 해당 수치가 증가할수록 향후 해당 지역의 재활용률 수치가 반대로 증가하는 현상이 발생함이 확인되었다. 이는 공업지역 규모가 일정 수준을 넘어서기 전까지는 폐기물 차원에서 향후 재활용을 저하하는 요소로 작용하지만, 공업지역 면적의 규모가 그 이상을 초과할 경우, 해당 단지의 큰 규모로 인해 폐기물 관리 감독 및 규제가 강하게 전개되며 역설적으로 향후 재활용률의 개선이 발생하게 되는 것이라 추측된다. 다음으로, 3인 이상 세대 수 PDP plot 분석 시 약 80,000세대까지는 해당 규모가 증가할수록 해당 자치구의 향후 재활용률이 증가하는 현상이 발생하나, 그 이후로는 규모의 증가가 오히려 재활용률의 악화에 기여하는 것으로 나타났다. 이는 선행연구에서 확인된 가구 형태와 재활용률 사이 인과구조가 (다가구 세대 기준) 약 8만 가구 수준까지는 강조되지만. 그 이후부터는 인구과밀 또는 인구 규모 성질 자체의 변화에 따라 상관관계가 뒤바뀔 가능성이 존재함을 내포한다고 말할 수 있다. 한편, 나머지 네 가지 주요 변수인 인당 상업지역 면적, 주거지역 면적, 위탁급식 영업체 수, 그리고 인당 녹지면적의 경우, 재활용률 수치와 매우 미약한 음의 상관관계를 유지, 즉 네 변수는 주변 효과 차원에서 재활용률에 유의한 영향을 미치지 않음이 확인되었다. 이에 해당 네 가지 변수의 경우, 그 자체로는 표면적으로 유의한 상관관계를 갖지 않을 수 있으나, 타 변수와의 상호작용을 통해 향후 재활용률 수치의 증감 상 중요 요인으로 작용한다는 것이 확인되었다.

〈그림 7〉 BART 모델 기준 변수 중요도 및 주요 변수(6개)별 주변효과 그래프



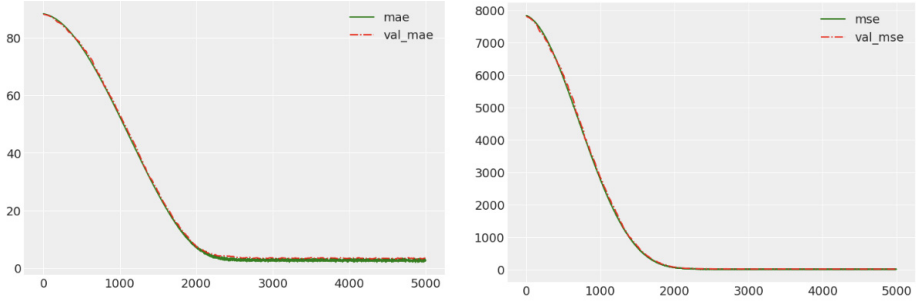
3. 심층 신경망(DNN) 학습 및 평가, 해석 결과

한편 5000번의 에포크를 전제로 DNN을 학습한 결과, 최적 하이퍼 파라미터 조합으로는 {Adam Compiler, Batch Size: 50}이 선택되었으며(표 2), 이때 최적 조합에 따른 모델 학습 결과 훈련 데이터에 대한 손실함수 및 평가도구 수치로서는 (완료 시점 기준) MSE: 8.75, MAE: 1.84%p가 도출됨을 확인하였다.[5] 또한, 해당 최적 조합과 관련하여 학습 데이터에 대한 에포크별 손실 함수값과 MAE를 시각화한 결과, (그림 8)과 같은 학습 곡선이 발생함을 확인하였다(성공적인 수렴 확인). 이에 해당 최적 조합을 전제로 한 신경망 모델을 기반으로 평가 데이터에 대한 재활용률 예측을 진행한 결과, MSE로는 9.79, MAE로는 2.28%p가 도출됨에 따라, DNN 모델 역시 앞선 BART 모형과 마찬가지로 우수한 설명력 및 예측력을 내포하고 있음을 최종적으로 확인하였다.

〈표 2〉 하이퍼 파라미터 조합별 훈련데이터 기준 MAE(MSE)

MAE(MSE)(%p)		Batch Size: 배치 크기(10,20,30,40,50)				
		10	20	30	40	50
Compiler	Adam	2.44(14.60)	2.31(12.69)	1.93(9.22)	2.25(13.31)	1.84(8.75)
	RMSprop	2.40(11.77)	2.28(11.55)	2.02(9.33)	2.20(11.07)	1.86(8.44)

〈그림 8〉 DNN 최적 모델 기준 학습 곡선(좌: MAE, 우: MSE)

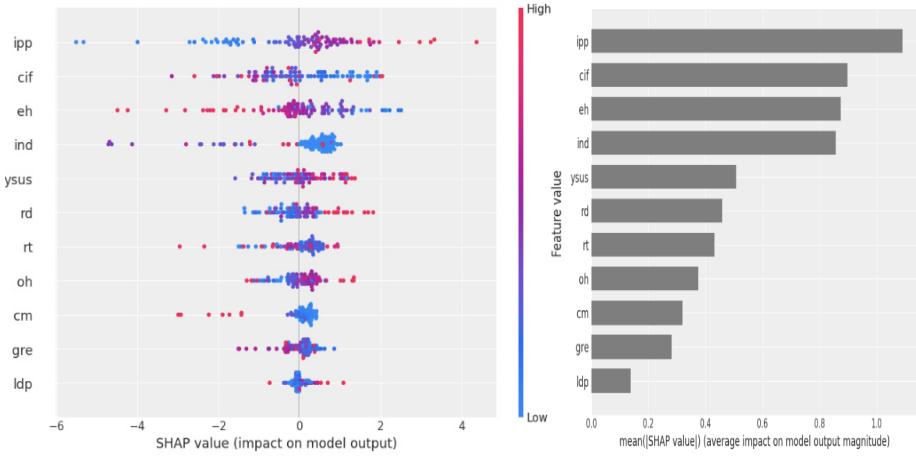


*val: validation(검증 데이터에 대한 MAE, MSE 표현)

5) 이때 표본 개수 상 한계가 존재하는 배치: 100을 제외한 모든 파라미터 조합에서는 학습 데이터의 10%를 별도 추출하여 검증 데이터를 구축하고, 이에 대한 손실 함수 값과 평가 도구상 수치를 함께 정리하여 모델별 성능을 비교하였다.

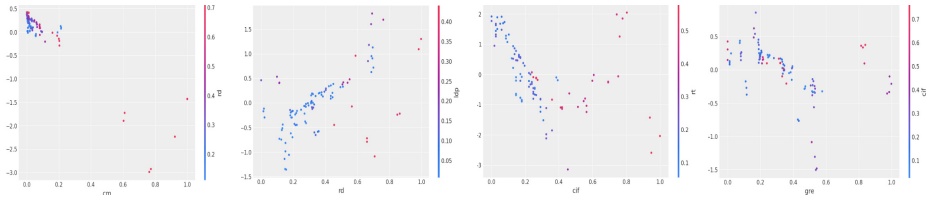
다음으로, 본 연구에서는 해당 DNN 모델의 성능이 유의함을 전제로 SHAP을 활용한 XAI 분석을 전개하였다. 이에 훈련 데이터를 전제로 shap 패키지의 KernelExplainer()를 적용한 결과 다음과 같은 변수별 중요도 및 한계 기여 방향성 그래프가 형성되었다(그림 9). 먼저 변수 중요도의 측면에서는 모델 학습 시 즉석판매제조가공업체 수, 위탁급식영업체 수, 3인 이상 세대 수, 인당 공업지역 면적, 유년부양비, 인당 주거지역 면적, 일반 음식점 수, 1인 세대 수, 인당 상업지역 면적, 인당 녹지지역 면적, 그리고 인당 지방세 부담액 순서로 향후 재활용률 증감 및 예측에 유의한 영향을 미치는 것으로 나타났다. 이때 변수 중요도의 크기 차원에서는 4번째 변수인 인당 공업지역 면적(ind) 변수를 기점으로 일종의 엘보우 포인트가 발생함이 확인되었으며, 이에 따라 본 DNN 모델 분석에서의 주요 변수로는 즉석판매제조가공업체 수(ipp), 위탁급식영업체 수(cif), 3인 이상 세대 수(eh), 그리고 인당 공업지역 면적(ind)을 선정하여 연구를 재개하였다. 변수 간 상호작용 전제 아래 주요 변수별 한계 기여도의 경우, 즉석 가공업체 수의 경우에는 양의 방향으로, 위탁급식영업체 수와 3인 이상 세대 수는 음의 방향으로 향후 재활용률에 영향을 미치는 것으로 나타났으며, 인당 공업지역 면적의 경우에는 영향의 방향성을 일반화하기에 무리가 있는 것으로 나타났다. 한편, Shapley 상호작용지수를 전제로 할 때 앞서 BART 모델에서 대두된 변수인 인당 상업지역 면적, 인당 주거지역 면적, 위탁급식영업체 수, 그리고 인당 녹지면적의 경우 SHAP 분석 시 주어진 변수집합 가운데 각각 주거지역 면적, 인당 지방세 부담액, 일반 음식점 수, 위탁

〈그림 9〉 Shapely value 기준 Summary Plot



급식영업체 수와 가장 강한 상호작용을 유지함이 확인되었으며(그림 10), 해당 상호작용 아래 네 변수는 한계 기여도 측면에서 향후 재활용률에 각각 음의 방향, 양의 방향, 음의 방향, 음의 방향으로 영향을 미치는 것으로 나타났다.

〈그림 10〉 Shapely value 기준 Dependence Plot.
(X축: 순차적으로 cm, rd, cif, gre 변수, Y축(좌): X축 변수에 대응되는 shapley 값, Y축(우): X축 변수에 대응되는 가장 강한 상호 작용 변수)



4. BART, DNN 모델 및 변수 분석 종합

BART 및 DNN 분석 결과, 자치구별 재활용률은 (시차를 두고 가정한 설명변수 집합을 전제로) 약 2%p의 평균 절댓값 오차 하에 유의하게 예측될 수 있음이 확인되었으며, 이때 변수 가운데 공통적으로 개인의 소득이나 원자재 가격, 또는 재활용품 자체의 가격 같은 경기 및 시장 관련 변수에 비하여 지역 내 상업·공업·주거 면적의 분포, 식품위생업종 현황, 그리고 세대별 가구원 수 등이 향후 재활용률의 증감에 상대적으로 중요한 영향을 미치는 것으로 나타났다. 구체적으로 BART 모델에서는 인당 공업지역 면적, 3인 이상 세대 수, 인당 상업지역 면적, 인당 주거지역 면적, 위탁급식영업체 수, 그리고 인당 녹지면적 변수가, DNN 모델에서는 즉석판매제조가공업체 수, 위탁급식영업체 수, 3인 이상 세대 수, 그리고 인당 공업지역 면적 변수가 핵심 변수로서 선정되었으며, 두 모델을 활용한 분석 아래 공통된 주요 변수로는 인당 공업지역 면적(ind), 위탁급식영업체 수(cif), 3인 이상 세대 수(eh)가 존재함이 확인되었다. 한편, BART를 활용한 분석에서는 인당 공업지역 면적과 3인 이상 세대 수의 주변 효과를 관찰 시 두 변수 모두에서 일정 수준의 초과가 발생할 경우 향후 재활용률과의 영향 관계 방향성이 역전되는 현상이 발견되었으며, DNN을 활용한 XAI 분석에서는 Shapley 상호작용지수 고려 시, 앞선 BART 분석에서 중요성은 높았으나 뚜렷한 주변 효과가 파악되지 않은 네 변수와 관련하여 각 변수가 주거지역 면적, 인당 지방세 부담액,

일반 음식점 수, 위탁급식영업체 수와 가장 강한 상호작용을 하며 향후 재활용률 증감에 영향을 미치는 현상이 대두되었다.

5. 재활용률 증진 정책 전개 시 BART 및 DNN 모델의 구체적 활용 방안

(예시: 2023년 자치구별 재활용률 예측을 기반으로 한 정책 제안)

앞선 분석을 종합하여, 본 연구에서는 선제적인 재활용률 증진 정책 전개 시 해당 학습 모델들을 활용할 실질적 방안을 제시하고자 하였으며, 이의 구체적 예시를 제공하고자 2023년 서울시 25개 자치구별 재활용률을 해당 모델들을 통해 선제적으로 예측하고, 앞선 주변효과 및 한계 기여도를 전제로 재활용률 악화 또는 취약 지역에 대한 대응 방안 제시를 진행하였다. 먼저 본 연구상 설명변수는 반응변수와 1 또는 2년 시차를 두고 설정되었으므로, 2021년 또는 2022년 설명변수 데이터셋 집합을 구축한 다음 앞선 과정과 동일하게 각 모델을 활용한 결과, (표 3)과 같은 재활용률 예측치가 도출되었다(본 연구에서는 예측되는 재활용에 따른 자치구 순위상 상위 또는 하위 20%에 집중. 이에 따라 25개 자치구 가운데 5개 자치구를 각 기준에 따른 핵심 자치구로 고려).

〈표 3〉 BART, DNN 모델별 2023년 자치구별 재활용률 예측(%)

BART 기반 예측치				DNN 기반 예측치			
자치구	예측치	자치구	예측치	자치구	예측치	자치구	예측치
종로구	86.12%	마포구	89.35%	종로구	87.81%	마포구	88.93%
중구	87.83%	양천구	84.08%	중구	84.34%	양천구	86.53%
용산구	86.88%	강서구	80.53%	용산구	89.79%	강서구	78.71%
성동구	82.25%	구로구	86.47%	성동구	82.1?%	구로구	86.35%
광진구	90.67%	금천구	87.26%	광진구	91.18%	금천구	88.91%
동대문구	90.43%	영등포구	88.09%	동대문구	90.8%	영등포구	91.05%
중랑구	92.49%	동작구	91.74%	중랑구	91.53%	동작구	91.8%
성북구	91.06%	관악구	91.38%	성북구	91.42%	관악구	90.74%
강북구	89.43%	서초구	89.32%	강북구	89.73%	서초구	91.47%
도봉구	81.81%	강남구	88.45%	도봉구	78.05%	강남구	87.17%
노원구	89.51%	송파구	88.32%	노원구	85.85%	송파구	90.55%
은평구	90.77%	강동구	92.42%	은평구	90.81%	강동구	89.92%
서대문구	89.22%			서대문구	90.37%		

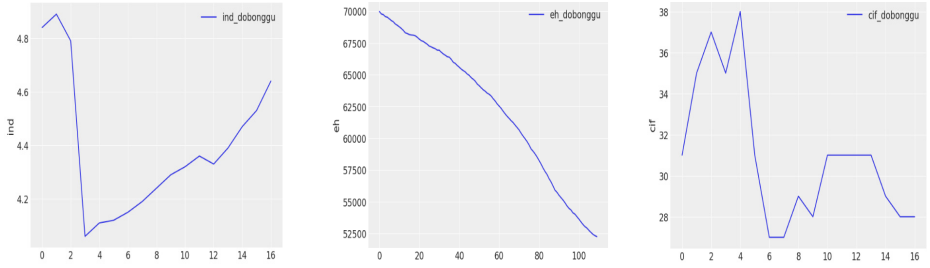
먼저 절대적 재활용률 수치와 관련하여서, BART 모형에서는 강서구, 도봉구, 성동구, 양천구, 종로구 순으로 재활용 수준이 낮을 것으로 예측되었으며, DNN 모델에서는 도봉구, 강서구, 성동구, 중구, 노원구 순으로 재활용 수준이 낮을 것으로 예측되었다. 다음으로, BART 모형에서는 코로나-19 펜데믹 시작 시점인 2020년과 대비 시 금천구, 도봉구, 동대문구, 송파구, 영등포구 순으로 2023년 재활용률 악화 정도가 높을 것으로 예측되었으며, DNN 모델에서는 도봉구, 금천구, 동대문구, 강남구, 중구 순으로 악화 정도가 높을 것으로 예측되었다. 이때 절대적 재활용률의 차원에서는 두 모델에서 공통적으로 강서구, 도봉구, 그리고 성동구가, 악화 정도의 차원에서는 공통적으로 금천구, 도봉구, 동대문구가 주요 자치구로 제시됨에 따라, 두 차원을 종합 시 선제적인 정책적 개입이 시급한 자치구로는 서울특별시 도봉구가 존재하는 것으로 나타났다.

이때 앞서 분석된 변수별 중요도 및 주변 효과와 한계 기여도를 기반으로 도출되는 공통 주요 변수를 전제로 할 시, 도봉구의 재활용률 증진을 위한 선제적인 정책적 개입 방안으로는 다음과 같은 제안이 이루어졌다.[6] 먼저 공통 중요 변수 가운데 하나인 인당 공업지역 변수의 경우, 2007년을 기점으로 급감한 후 점진적으로 증가하며 기존 수준을 회복함으로써 2021년 기준 공업지역 면적은 인당 $4.64m^2$ 수준인 것으로 나타났다. 즉, BART 변수 분석에서 확인되는 인당 공업지역 면적의 주변 효과를 고려 시, 2023년 도봉구에서 예측되는 재활용률 악화는 지역 내 공업지역 면적이 $5m^2$를 초과하지 않는 선에서 상당 수준 증가하는 추세가 나타남에 따라 발생했다고 해석할 수 있다. 이에 따라 정책적 차원에서는 도봉구 내 공업지역 면적의 축소(공업 단지 분산 또는 이전)를 촉진하거나, 현재 존재하는 공업지역 내 폐기물 처리 관리 감독을 강화하는 동시에 재활용 인센티브를 지급하는 방식으로 개입을 전개함으로써, 예측되는 재활용 악화를 부분적으로 해결할 수 있다고 파악된다.

---

6) 이때 추세분석 대상이 되는 주요 변수별(도봉구 인당 공업지역 변수, 3인 이상 세대 수, 그리고 위탁급식영업체 수 변수) 시계열로는 접근 가능한 데이터의 범위 등을 고려하여 각각 2005년~2021년(연도별), 2013년 11월~2022년 12월(월별), 2005년~2021년(연도별) 시계열 자료를 활용

〈그림 11〉 중요 영향 변수별 시계열: 좌측부터 차례로 도봉구의 인당 공업지역면적, 3인 이상 세대 수, 위탁급식영업체 수



주요 변수 가운데 3인 이상 세대 수의 경우에는, 시계열상 2013년부터 2022년 현재까지 지속적인 감소 추세(69979 세대에서 52252 세대로 감소)가 확인되었다. 즉, 현재 도봉구에서 예측되는 재활용률의 악화는 (BART 변수 분석상 확인되었던 영향 관계 고려 시) 8만 세대 아래에서 도봉구 내 3인 이상 세대 수가 지속적으로 감소하는 추세가 전개됨에 따라 강화되었다고 해석할 수 있다. 이에 따라 정책적 개입의 차원에서는 도봉구 내 교육기관 및 돌봄 기관 등을 확대함으로써 자녀 양육 가구 유입을 확대하는 방식의 정책을 전개하거나, 여러 사람이 한 세대를 공유하는 형태의 셰어하우스 단지 등을 확대 조성 및 추진하려는 노력이 예측되는 재활용률 개선에 부분적으로 기여할 수 있을 것이라 파악된다.

한편, 주요 변수 가운데 도봉구의 위탁급식영업체 수는 시계열 상 2009년을 기점으로 급감한 다음 2021년까지는 2005년 수준(31개) 이하에서 진동하는 추세가 확인되었다. 이는 DNN에서 확인되는 한계기여도 특성, 즉 일반 음식점 수와의 상호작용 아래 향후 재활용률에 음의 방향으로의 영향을 미치는 특성을 고려 시, 상호작용 효과가 현재 도봉구에서 예측되는 재활용률 상 악화에 유의미하게 기여하지 않았음을 의미하며, 이에 따라 정책적 차원에서 위탁급식영업체의 폐기물 배출에 대한 정책적 개입 등을 진행할 필요성은 적은 것으로 나타났다. 따라서 2023년에 예측되는 도봉구 재활용률 악화를 해결하기 위한 선제적 개입으로는 주요 변수 가운데 도시지역 면적(분포)과 가구 형태에 대한 정책을 전개하는 것이 효과적일 것이며, 구체적인 방안으로는 도봉구의 공업지역 면적 축소 또는 공업 지역에서의 폐기물 배출 모니터링 강화 정책과 함께 양육 환경 개선 또는 셰어하우스 단지 조성을 통한 3인 이상 세대 유입 증진 정책을 전개하는 것이 효과적일 것이라 제안할 수 있다고 판단된다.

이처럼 1년 내지 2년 이후 특정 자치구에서 발생할 재활용률의 개선 또는 악화는 BART와 DNN 모델을 통해 유의하게 예측할 수 있음이 확인되었으며, 해당 예측치에 따른 향후 특정 지역 내 재활용에 대한 정책적 대응은 (위 2023년 예시와 같이) 변수별 시계열 추세분석과 함께 앞서 두 모델을 기반으로 확인된 변수별 주변 효과 및 한계 기여도의 방향성 및 특성을 종합하는 방식으로 직접적 원인을 탐지하는 동시에 구체적 해결 방안을 제시함으로써 효과적으로 전개될 수 있을 것이라 예상된다.

V. 결론 및 제언

본 연구는 현 사회가 코로나-19 펜데믹을 거치며 직면하게 된 폐기물 차원상 탄소 중립 저해 현상의 심각성에 집중하여, 이를 완화하기 위한 선제적 방안을 모색하는 것을 주요 연구 목적으로 설정하였다. 이에 환경부와 환경관리공단에서 제공하는 전국 폐기물 통계, 통계청 KOSIS e-지방지표 등을 기반으로 자치구별 재활용률을 반응변수로, 1년, 내지 2년 시차를 둔 자치구별 인구구조, 거주형태, 식품위생업 현황, 도시면적 분포 데이터 등을 설명변수로 설정하여 홀드 아웃 샘플링을 진행한 다음, 훈련 데이터 및 평가 데이터를 기반으로 베이지안 가법 회귀 모형(BART)과 심층 신경망(DNN)을 학습 및 평가하였다. 이에 따라 학습된 각 모델을 전제로 특정 지역에서 발생할 재활용률을 약 1년 시차를 두고 예측하는 데 기여하는 주요 변수 및 변수별 주변 효과와 한계기여도를 추출, 이와 함께 예측 진행 및 이에 따른 재활용률 증진을 위한 실질적인 정책적 개입 시 해당 모델들의 구체적 활용 방안을 제시하였다.

연구 결과는 다음과 같다. 먼저 BART 모형과 DNN 모형에서 홀드아웃 샘플링을 기반으로 훈련 데이터에 대한 학습 및 평가 데이터에 대한 평균 제곱 오차(MAE)를 확인한 결과, BART 모형에서는 전체 훈련 데이터의 4% 정도의 표본을 제외한 모든 데이터에서 MCMC 체인의 수렴이 확인되는 동시에 평가 데이터에 대하여 약 2.66%p의 MAE가 발생하였으며, DNN 모델에서는 평가 데이터에 대하여 약 2.28%p의 MAE가 도출됨에 따라 두 모델 모두에서 매우 적은 오차를 두고 향후 재활용률을 유의하게 예측할 수 있음이 나타났으며, 이에 폐기물 처리상 탄소 중립을 실현하는 데 두 모델이 유의미한 역할을 수행할 수 있음이 확인되었다.

다음으로, BART 모델 분석상 변수별 중요도를 확인한 결과, 전체 설명변수 집합 가운데 자치구별 1인당 공업지역 면적, 3인 이상 세대 수, 1인당 상업지역 면적, 1인당 주거지역 면적, 위탁급식영업체 수, 그리고 1인당 녹지면적이 주요 변수로 선정되었으며, 이때 주변 효과상 지역 내 1인당 공업지역 면적은 약 $5m^2$을 기점으로 그 이전까지는 향후 재활용률에 대하여 음의 방향으로 영향을, 그 이후부터는 양의 방향으로 영향을 미치는 것으로 나타났으며, 지역 내 3인 이상 세대 수는 8만 가구를 기점으로 하여 그 이전까지는 향후 재활용률 증진에 기여하나 그 이후로는 음의 방향으로 영향을 미치는 것으로 나타났다. 이는 공업지역 규모의 경우, 일정 수준을 넘어서기 전까지는 폐기물 차원에서 향후 재활용을 저하하는 요소로 작용하지만, 공업지역 면적의 규모가 그 이상을 초과할 경우, 해당 단지의 큰 규모로 인해 폐기물 관리 감독 및 규제가 강하게 전개됨으로써 통념과는 배치되는 결과가 발생하는 것으로 해석할 수 있으며, 3인 이상 세대 수의 경우에는 선행연구에서 확인된 가구 형태와 재활용률 사이 인과구조가 (다가구 세대 기준) 약 8만 가구 수준까지는 강조되지만. 그 이후부터는 인구과밀 또는 인구 규모 성질 자체의 변화에 따라 상관관계가 뒤바뀔 가능성이 존재하는 것으로 해석 가능하다.

한편, SHAP을 기반으로 한 DNN 모델 분석상 변수별 중요도를 확인한 결과, 전체 설명변수 집합 가운데 즉석판매제조가공업체 수, 위탁급식영업체 수, 3인 이상 세대 수, 그리고 인당 공업지역 면적이 주요 변수로 선정되었으며, 이때 상호작용 전제 하에 각 변수별 한계 기여도를 고려 시 즉석 가공업체 수의 경우에는 양의 방향으로, 위탁급식영업체 수와 3인 이상 세대 수는 음의 방향으로 향후 재활용률에 기여하는 것으로 나타났으며, 인당 공업지역 면적의 한계 기여 방향성은 일반화하기에 어려움이 존재하는 것으로 나타났다. 또한, Shapley 상호작용 지수 전제하 변수 간 의존도를 분석한 결과, 앞서 BART에서 추출된 주요 변수 가운데 특별한 주변 효과가 파악되지 않았던 자치구별 1인당 상업지역 면적, 주거지역 면적, 위탁급식영업체 수, 그리고 녹지지역 면적은 각각 주거지역 면적, 인당 지방세 부담액, 일반 음식점 수, 위탁급식영업체 수와 가장 강한 상호작용을 유지하며 각각 향후 재활용률에 각각 음의 방향, 양의 방향, 음의 방향, 음의 방향으로 한계 기여하는 것으로 나타났다.

마지막으로, 본 연구에서 사용된 BART 모형과 DNN 모델의 구체적 활용 방안 분석 결과는 다음과 같다. 본 연구에서는 앞서 추출된 변수별 주변효과 및

한계 기여도와 각 변수별 시계열 추세를 통해 특정 지역에서의 재활용률 개선 또는 악화 원인 및 대응방안을 추출할 수 있다고 판단하였으며, 이때 구체적 예시로서 2023년 서울시 25개 자치구별 재활용률을 예측, 그 결과 서울특별시 도봉구의 경우 절대적 수치와 상대적 변화 모두에서 가장 강한 재활용률 악화가 발생할 것으로 예측되었다. 이에 따라 두 모델에서 공통적으로 강조된 주요 변수인 인당 공업지역 면적, 3인 이상 세대 수, 그리고 위탁급식영업체 수를 정책적 개입 변수로 설정한 다음, 각 변수별 시계열과 영향 관계를 종합적으로 연계한 결과 세 요인 가운데 인당 공업지역 면적과 3인 이상 세대 수가 해당 지역의 재활용률 악화에 원인으로 작용하였다고 판단할 수 있음이 나타났으며, 이에 대한 선제적 개선 방안으로는 공업지역 면적 축소 또는 공업단지에서의 폐기물 배출 모니터링 강화 정책과 함께 양육 환경 개선 또는 셰어하우스 단지 조성을 통한 3인 이상 세대 유입 증진 정책 등을 가정할 수 있다고 판단되었다.

본 연구의 의의 및 시사점은 다음과 같다. 첫째, 본 연구는 전통적인 통계 모델이 아닌 머신러닝 및 베이지안 MCMC 샘플링을 기반으로 한 BART와 딥러닝을 전제로 한 DNN을 활용하여 선제적으로 재활용률을 매우 적은 오차를 갖고 예측할 수 있도록 함으로써 탄소 중립의 정책 수립상 실질적인 방향성 수립에 기여할 수 있다는 점에서 의의를 갖는다. 둘째, 본 연구는 동일 시점에서의 재활용률 영향 요인을 추출했던 기존 연구와는 달리 선제적 개입 가능성, 데이터의 수집기간, 공시 기간 등을 고려하여 1년 이상의 시차를 둔 설명변수 집합을 구성함으로써 실질적 예측 가능성을 높였으며, 이때 각 변수별 주변효과와 한계기여도를 동시에 고려함으로써 변수별 주효과와 교호작용 효과를 함께 제시했다는 점에서 또한 의의를 갖는다. 셋째, 본 연구는 앞선 두 모델의 학습 및 분석을 넘어 정책 수립 시 해당 모델들의 실질적 활용 방안을 제시하기 위해 2021, 2022년 데이터를 기반으로 2023년 서울특별시 자치구별 재활용률 예측을 전개함으로써 실제적 선례를 제공한다는 점에서도 큰 의의를 갖는다.

본 연구의 한계점은 다음과 같다. 첫째, 분석의 범위가 전국 단위가 아닌 서울특별시였다는 점을 고려 시, 기타 지역의 설명변수 특성과 재활용률 사이의 연관은 본 연구에서 확인된 관계와는 다소 차이가 존재할 수 있다는 측면에서 한계가 존재한다. 단적으로 농업, 임업, 어업과 같은 1차 산업 중심 지역의 경우, 재활용률 예측상 본 연구에서 강조되었던 공업지역, 상업지역 면적 등의 중요도가 매우 낮을

수밖에 없으며 고령화 및 적은 인구 유입 등으로 인해 인구구조의 영향이 대도시에 비하여 상대적으로 강조된다는 점에서 본 연구에서 분석된 재활용률 영향 관계와 해당 지역에서의 영향 관계 양상은 다소 차이가 존재할 수밖에 없다고 판단된다. 이에 후속 연구 등이 진행될 시, 전국 단위, 또는 도심 외 지역을 전제로 지역적 특성과 재활용률 사이의 관계가 재분석될 필요성이 존재한다고 사료된다.

둘째, 본 연구는 자치구 단위의 접근을 취함으로써 해당 지역 내 재활용률 악화 또는 개선 수준을 분석한다는 점에서, 다소 큰 지역 구분을 통해 일반화를 전개한다는 단점을 갖는다. 즉, 본 연구는 행정동, 또는 그보다도 더 세분화된 지역 구분을 전개할 시 발생하는 미시적인 수준의 재활용 영향 요인들을 고려하지 못한다는 측면에서 분석된 변수 집합이 거시적 분석에 다소 치중되었다고 말할 수 있다. 따라서 관련 후속 연구 또는 연관 분석에서는 본 연구가 다루지 못한 미시적인 폐기물 배출 행태 영향요인 및 해당 요인이 재활용률에 미치는 영향을 추가적으로 분석함으로써, 본 연구에서 구축된 변수 탐색의 지평을 한층 더 넓힐 필요성이 존재한다고 사료된다.

Ⅵ. 참고문헌

김지욱, 정의철(1998), "가계의 쓰레기 재활용시간 결정요인 분석," 『경제학연구』 46(1): 255~271.

서울 열린데이터광장, 서울시 폐기물 재활용 통계, 서울특별시 기본통계

서울 열린데이터광장, 서울시 부양비 및 노령화지수 (구별) 통계, 주민등록인구통계

서울 열린데이터광장, 서울시 세대원수별 세대수(구별) 통계, 주민등록인구 통계

서울 열린데이터광장, 서울시 식품위생업 현황(구별) 통계, 서울특별시 기본통계

서울 열린데이터광장, 서울시 지방세 부담 통계, 서울특별시 기본통계

양진우, 박해식(2003), “경로분석을 이용한 생활폐기물 분리배출 및 재활용 행동의 영향요인에 관한 인과구조분석,” 『국토계획』, 38(3): 233~244.

유두련(2002), “재활용 행동 집단별 소비자특성과 영향요인에 관한 비교연구,” 『대한가정학회지』, 40(6): 53~67.

이소라(2018), “재활용산업 활성화를 위한 지원 방안의 영향분석,” 『환경정책』, 26(2): 167~190.

이재준 외 3인(2021), "XGBoost와 SHAP 기법을 활용한 근로자 이직 예측에 관한 연구," 『정보시스템연구』, 30(4): 21~42.

유광민, 박정원(2022), "생활계폐기물 및 재활용품 배출에 미치는 영향요인 분석: 종량제 정책 수단을 중심으로," 『지방행정연구』, 36(4): 337~367.

질병관리청, COVID-19 DashBoard

추한경 외 4인(2018), "점진적 샘플링과 정규 상호정보량을 이용한 온라인 기계학습 공조기 급기온도 예측 모델 개발," 『대한건축학회 논문집-구조계』, 34(6): 63~69.

통계청, 『온라인쇼핑동향조사』, 취급상품범위별/상품군별거래액

통계청, KOSIS, e-지방지표, 1인당 도시지역면적 현황(시도/시/군/구), /도시계획현황/, 한국국토정보공사

통계청(물가동향과), KOSIS, e-지방지표, 소비자물가 등락률(시도/시)

한준(2020), "국내 생활폐기물 분야 플라스틱 비재활용 처리량 요인분해 연구," 『환경정책』, 28(2): 79~100.

홍성훈(2015), "종량제 가격이 생활폐기물, 음식물쓰레기, 재활용품 수거서비스 수요에 미치는 영향," 『자원·환경경제연구』, 24(4): 747~761.

현승현, 정지훈(2017), "지방정부의 민간위탁 수준이 생활폐기물 처리에 미치는 영향요인 분석: 경기도 기초자치단체를 중심으로," 『지방행정연구』, 31(4): 177~198.

환경부, 환경통계포털〉 자원순환〉 재활용가능자원 가격조사〉 원자재 수입동향〉 원자재 수입동향 종합

환경부, 환경통계포털〉 자원순환〉 재활용가능자원 가격조사〉 재활용 품목별 가격동향〉 지역 및 품목별 가격현황

Chipman H.A. et al.(2010), "BART: BAYESIAN ADDITIVE REGRESSION TREES," The Annals of Applied Statistics, 4(1): 266~298.

Dickinson, Quinn; Meyer, Jesse G.(2022), "Positional SHAP(PoSHAP) for Interpretation of machine learning models trained from biological sequences," ALoS computational biology, 18(1)

Glorot X, Bengio Y (2010), "Understanding the difficulty of training deep feedforward neural networks," In: Proceedings of the thirteenth

international conference on artificial intelligence and statistics, 249~256.

Gunning D., et al.(2019),"XAI-Explainable artificial intelligence," Science robotis, 37(4)

He et al(2015), “Delving deep into rectifiers: surpassing human-level performance on imagenet classification,” In: Proceedings of the IEEE international conference on computer vision, 1026~1034.

Kook Hyun Yoo(2018), "Weight Initialization of neural network by using independent component anlaysis and restricted random noises," Ph.D., Hanyang University. 1~66.

Lundberg, Scott M., Gabriel G.Erion, and Su-In Lee(2018), "Consistent individualied feature attribution for tree ensembles," arXiv:1802.03888.

Narkhede, M.V. et al.(2022), "A review on weight initialization strategies for neural networks," The Artificial Intelligence Review, 55(1): 291~322.

Rothman, Denis. (2022), Hands-On Explainable AI(XAI) with Python, DK ROAD BOOKS

Shannon, C. (1948), "A Mathematical Theory of Communication,“ Bell System Technical Journal, 27(3): :379~423.

Yoon Y.A., Lee S.H., Kim Y.S.(2021), "A Study on the Remaining Useful Life Prediction Performance Variation based on Identification and Selection by using SHAP,“ Journal of Korean Society of Industrial Systems Engineering, 44(4): 1~11.