



장려상

머신러닝을 활용한 임차인 주거불안 예측모형 개발 및 실증

- 수도권 1인 가구를 중심으로 -

| 구한민 |

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



머신러닝을 활용한 임차인 주거불안 예측모형 개발 및 실증

- 수도권 1인 가구를 중심으로 -

Predictive Modeling of Housing Instability among One-Person
Tenant Household in the Seoul Metropolitan Area, South Korea:
An Ensemble-Based Machine Learning Approach

구한민*

요약

최근 수도권 1인 가구의 주거 상황이 취약하므로 정책적으로 선제적인 조치가 필요한 상황이다. 본 연구에서는 앙상블 기반의 머신러닝을 활용하여 수도권 1인 가구 임차인의 주거불안 예측모형을 개발하였다. 모델링 결과, 랜덤 포레스트로 구축한 모형이 가장 우수한 예측성능을 나타냈다. 이를 실제 정책과정에 활용한다면 임차인이 주거위기의 상황에 빠지기 전 예방적 조치를 취할 수 있을 것이다. 한편 이들이 주관적으로 경험하고 느끼는 주거불안은 객관적인 주거불안정의 지표와 괴리될 수 있음을 확인하였다. 따라서 주거 복지정책은 현행대로 최저주거기준, 주거비 등 객관적인 지표에 기반하되, 주관적 경험에 대한 보고는 보조적 지표로 활용하는 것이 바람직하다. 아울러 객관적인 상황과 괴리된 판단으로 불안을 겪는 가구를 위하여 주거 관련 상담서비스를 제공할 필요가 있다.

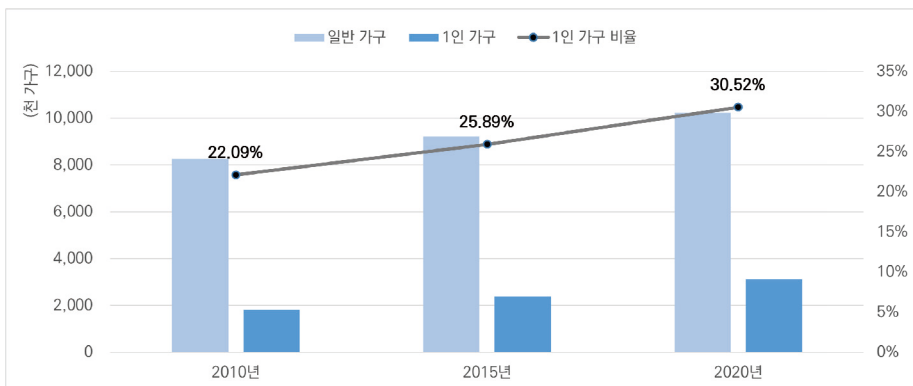
주제어: 주거불안, 예측모형 개발, 수도권 1인 가구, 머신러닝, 앙상블 기법

* 연세대학교 도시공학과 석박사통합과정

I. 서 론

최근 우리나라의 주요한 인구 이슈는 수도권 인구집중과 1인 가구의 증가다. 2021년 11월 기준, 총인구 5,174만 명 중 서울·인천·경기 등 수도권 인구는 2,608만 명으로 50%에 육박한다(통계청, 2022a). 한편 우리나라 총인구는 감소세로 돌아섰음에도 급격한 가구 분화로 가구수는 오히려 증가하고 있다. 특히 2000년대를 기점으로 급증한 1인 가구는 2021년 기준 약 720만 가구로 전체 가구에서 33.4%를 차지하고 있다(통계청, 2022a). 통계청(2022d)의 추계에 따르면 2050년 1인 가구는 905만 가구로 늘어나 전체 가구의 약 40%를 차지하는 주된 가구 유형이 될 것으로 보인다. 수도권 인구집중과 1인 가구의 증가라는 인구 트렌드는 수도권 1인 가구의 급증 현상으로 귀결되고 있다(〈그림 1〉). 2021년 현재 전국 1인 가구 중 약 47%(서울 20.8%, 경기 21.5%, 인천 5.0%)를 차지하는 수도권의 1인 가구(통계청, 2022b)는 머지않아 정부의 주요한 정책 대상이 될 것으로 보인다. 따라서 이들에게 필요한 정책을 선제적으로 수립할 필요가 있다.

〈그림 1〉 2010~2020년 수도권 1인 가구의 변화(통계청(2022c) 바탕으로
저자 작성)



특히 정책 입안자는 수도권 1인 가구의 취약한 주거 상황에 초점을 맞추어야 한다. 김옥연·문영기(2009), 반정호(2012), 권연화·최열(2018) 등 다수의 실증 연구에 따르면 우리나라의 1인 가구는 청년, 중·장년, 노년 등 연령대를 가리지 않고 사회적 취약계층일 가능성이 높았다. 또한 이들은 다양한 측면에서 주거의

불안을 경험하고 있었다(이동훈, 2012; 강은나·이민홍, 2016; 박미선 외, 2017; 정인·오상엽, 2020; 구한민·김갑성, 2023). 이에 수도권 소재의 서울연구원(변미리 외, 2015), 인천연구원(김윤영·김아진, 2017), 경기연구원(최석현·김재신, 2017) 등에서는 각 지자체의 1인 가구 특성을 분석하고 그에 따른 주거정책을 제안한 바 있다. 하지만 실질적으로 하나의 도시권역으로 기능하는 수도권에 대한 통합적인 접근이 이루어지지 않았다는 점에서 이들 공간을 아우르는 연구의 필요성이 제기된다.

이러한 배경에서는 수도권 1인 가구의 주거불안의 원인을 규명하려는 인과적 접근보다 주거불안을 선제적으로 예측할 수 있는 모형을 개발하고 검증하는 등의 데이터 중심 접근법(data-driven approach)이 유용할 수 있다. 이 과정을 통하여 누가 주거불안을 경험하고 있는가, 경험하게 될 것인가를 스크리닝(screening)할 수 있기 때문이다. 이는 정책 대상에 대한 사전적 조치를 가능하게 한다는 목표하에서는 주거불안이 어떠한 메커니즘으로 형성되었는가를 사후적으로 설명하는 것보다 더욱 적합한 방법이다. 아울러 개발한 모형으로 주거복지의 수요자와 수요량을 예측하면 정책의 효율성을 높일 수도 있을 것으로 판단된다.

본 연구의 핵심 목적은 머신러닝(machine learning)을 활용하여 수도권 1인 가구 임차인의 주거불안 예측모형을 개발하고 검증하는 것이다. 모형으로는 예측 성능이 우수한 것으로 알려진 앙상블 기법(ensemble methods)을 적용하였고, 모델링(modeling)을 위한 데이터세트는 『2019년도 주거실태조사』의 응답으로 구축하였다. 앙상블 기반의 머신러닝은 표본 또는 관측치의 어떠한 특성이 주거불안 경험의 여부를 구분하는 데 상대적으로 더 큰 영향을 미치는지 중요성을 식별할 수 있게 하므로 이에 따른 추가적인 정책적 함의 역시 기대할 수 있다.

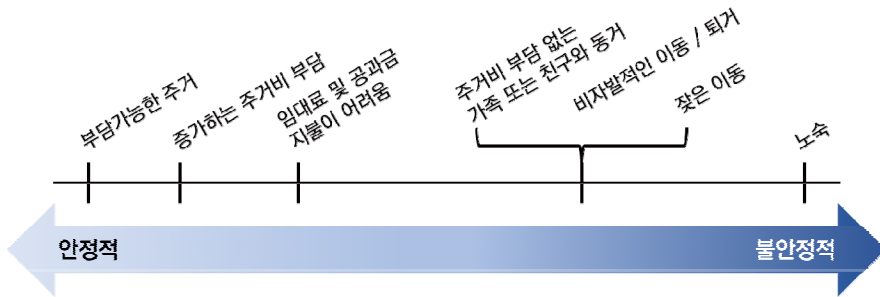
II. 문헌검토

1. 1인 가구의 주거불안

실증분석에 앞서 본 연구에서의 핵심 개념인 주거불안을 명확하게 정의할 필요가 있다. 추상의 구체화를 위하여 주거불안정(housing instability)의 개념에서 시작한다. 주거불안정의 정의는 합일되어있지 않으나(O'Campo et al., 2016),

넓은 의미에서는 “어떤 가구가 스스로의 주거를 충분히 통제할 수 없는 여러 상황”을 의미한다(Kang, 2019). 구체적으로 주거비에 대한 부담을 느끼는 상황, 과밀하게 주거하거나 가족 또는 친구와 동거하는 상황, 원치 않게 잦은 이동을 하는 상황, 퇴거 위기 또는 노숙자(homeless)의 지위에 놓이는 상황 등을 포함한다(Pavao et al., 2007; Rollins et al., 2012; Kang, 2019). 이러한 상황들은 <그림 2>와 같이 연속적, 누적적 영향을 미치며 주거불안정 연속체(housing instability continuum)¹⁾를 형성한다(Kleit et al., 2016).

〈그림 2〉 저소득층 가구의 주거불안정 연속체
(Kleit et al. (2016)의 그림을 재작성)



본 연구에서는 주거비 과부담, 과밀주거, 최저주거기준 미달 등 객관적인 주거불안정의 상황이 아닌 임차인이 느끼는 주관적인 주거불안정의 경험에 주목한다. 이는 대체로 임대인이 우월적인 지위에서 임차인과 임대차 관계를 형성하는 우리나라 임대주택시장의 특성을 고려하면(구한민·김갑성, 2021), 수치화한 지표만으로는 포착하기 어려운 주거불안정에 대한 실체적 접근에 가까울 수 있기 때문이다. 따라서 본 연구에서의 주거불안은 주거와 관련한 불안정성(precariousness)보다 불안감(anxiety)에 가깝다.

한편 세계적으로 개인주의가 만연하고 가족 내 소속감과 유대감 약화에 따른 핵가족화가 가속하면서 가구 문화가 사회적 현상으로 대두하기 시작되었다(Kobrin, 1976). 이에 따라 자연스레 1인 가구가 증가하게 되었다(Belcher, 1967). 우리나라 역시 1975년 최초로 통계를 집계한 이래 1인 가구가 꾸준히 증가하였으며,

1) 이를 제안한 Kleit et al. (2016)에 따르면 주거불안정의 심화 과정이 아니라 심각성을 나타내는 것이므로 반드시 한 방향으로만 이동하는 것은 아니다.

2000년대 급속히 증가하여 2021년 기준 전체 가구에서 약 33%를 차지하게 되었다(통계청, 2022a). 1인 가구 형성의 주원인은 연령대에 따라 청년은 미혼·만혼, 장년층은 이혼, 노년층은 사별 등으로 달랐으나(신상영, 2010; 김재익, 2013), 결과적으로 경제적으로 취약한 저소득층이 될 가능성이 높았다(김옥연·문영기, 2009; 반정호, 2012; 권연화·최열, 2018). 최근 통계청(2022b; 2022c)의 1인 가구 통계 역시 이를 방증한다. 2018년 기준 1인 가구의 연소득은 2,116만 원으로 전체 가구의 36.3% 수준에 그쳤지만, 총지출 중 임대료와 수도 광열비 등 주거비성 지출의 비율은 17.9%로 전체 가구 11.3%와 비교하여 크게 웃돌았다. 따라서 본 연구는 1인 가구 임차인이 경험하는 주거불안에 초점을 맞추었다.

2. 주거 연구와 머신러닝

최근 기술의 발전이 가속화하면서 복잡계(complex system)인 도시의 현상을 더욱 체계적으로 이해하기 위한 도구로서 빅데이터와 머신러닝의 활용이 확대되고 있다(정지은, 2022; 구한민·홍사흠, 2023). 특히 주거 분야의 연구자들은 주로 예측 분석에 머신러닝을 이용하였다. 대부분의 연구에서는 지도학습(supervised learning)²⁾으로 주택가격 결정요인을 분석하거나 미래 시점의 주택가격을 추정하였다(Fan et al., 2006; 이준용 외, 2007; 송호창 외, 2008; Özsoy & Şahin, 2009; 이창로, 2015; 배성완·유정석, 2018; 나성호·김종우, 2019; 홍정의, 2021; Gupta et al., 2021). 분석 알고리즘으로는 의사결정나무(decision tree), 랜덤 포레스트(random forest), 서포트벡터머신(support vector machine, SVM) 등이 적용되었다. 손철(2011), 이석준(2019) 등 일부 연구에서는 군집분석 등의 비지도 학습(unsupervised learning)³⁾을 통하여 주택하위시장(housing submarkets)을 구분하고 식별하였다.

근래에는 머신러닝을 적용한 연구의 주제가 더욱 다양해지고 있다. 대부분 의사결정나무에 기반한 알고리즘을 활용하였고 서울시 4개 구의 공동주택 공용

2) 예측변수를 바탕으로 결과변수를 추정하는 통계적 모형을 만드는 머신러닝 기법이다 (James et al., 2013).

3) 예측변수만 존재하여 이들의 특성을 추출하거나 패턴을 발견하는 머신러닝 기법이다 (Hastie et al., 2009).

관리비 추산(정진호 외, 2017), 수도권 지역의 주거이동 패턴 식별(Yi & Kim, 2018), 전국의 주택매도 결정요인 분석(김은미 외, 2020) 등을 시도하였다. 공유주택의 확산이라는 사회적 트렌드(김정기 외, 2023)에 알맞게 인공지능경망으로 공유주택 거주 의사를 분석한 연구(정기성, 2021)도 있었다.

〈표 1〉 머신러닝을 활용한 주거 분야의 선행연구

저자	연구목적	알고리즘
Fan et al. (2006)	주택가격 결정요인 분석	의사결정나무
이준용 외(2007)	주택가격 예측	의사결정나무, 인공신경망 등
송호창 외(2008)	주택가격 결정요인 분석	의사결정나무
Özsoy & Şahin (2009)	주택가격 결정요인 분석	의사결정나무
손철(2011)	주택하위시장 식별	군집분석
이창로(2015)	주택가격 예측	RF, 부스팅, SVM 등
정진호 외(2017)	공동주택 공용관리비 추산	RF
배성완·유정석(2018)	주택가격 추정	RF, 부스팅, 심층신경망 등
Yi & Kim (2018)	주거이동 패턴 식별	의사결정나무
나성호·김종우(2019)	주택가격 예측	RF, SVM 등
이석준(2019)	주택하위시장 구분	군집분석
김은미 외(2020)	주택매도 결정요인 분석	의사결정나무, RF, SVM 등
정기성(2021)	공유주택 거주의사 분석	인공신경망
홍정의(2021)	주택가격 추정	RF
Gupta et al. (2021)	주택가격 예측	RF

주: RF=random forest, SVM=support vector machine

저자가 인지하는 한 본 연구는 머신러닝을 적용하여 주거불안에 접근하는 초기적 시도이다. 이러한 접근의 필요성은 다음과 같다. 첫째, 수도권 1인 가구의 주거 취약성을 고려한 선제적 대책이 요구되는 현시점에서, 머신러닝을 활용하여 개발한 주거불안 예측모형은 복지정책 수요자의 스크리닝을 용이하게 만든다. 둘째, 앙상블 기반의 머신러닝은 로지스틱 회귀모형 등 전통적인 기법과 비교하여 가구의 주거불안 경험 여부에 영향을 미치는 특성을 더욱 잘 식별하므로, 주거정책이 어떠한 측면에 초점을 맞추어야 하는지 제언할 수 있다.

Ⅲ. 연구방법

1. 데이터세트

모델링을 위한 데이터세트의 구성에는 『2019년도 주거실태조사』의 응답을 활용하였다. 『2019년도 주거실태조사』는 국토교통부가 국민의 주거생활에 관한 현황 파악을 목적으로 2019년 하반기에 수행하였다. 본 연구의 대상적 범위는 서울·인천·경기 등 수도권에 거주하는 1인 가구 임차인이므로 전체 표본에서 이에 알맞은 케이스를 추출하였다. 연구의 목적에 알맞게 현재 거주 중인 임대주택 외에 가구주 또는 가구원의 이름으로 자가를 보유하고 있는 가구는 제외하였다. 또한 보증금, 월세, 소득, 자산, 부채 등 민감한 개인정보에 응답하지 않아 발생한 결측값은 목록별 삭제(listwise deletion) 방식으로 제거하였다. 전처리 결과 최종적으로 2,892가구의 응답을 데이터세트로 구축하였다.

2. 변수

1) 결과변수: 주거불안

결과변수는 임차인으로서의 주거불안(*ANXI*)을 현재 경험하고 있는지 여부로 0='불안하지 않다', 1='불안하다'의 이항형으로 측정하였다. 구체적으로 주거불안 경험 여부는 계약기간 중 임대인이 나가달라고 요구하는 경우에 대한 불안감, 계약기간 만료 후 임대인의 재계약 거부에 대한 불안감, 재계약 시 상승할 임대료 또는 전세의 월세 전환에 대한 불안감, 임대보증금을 돌려받지 못할 수도 있다는 불안감 중 하나라도 느끼고 있는 경우 주거불안을 경험하고 있는 것으로 파악하였다.

2) 예측변수: 가구 및 주택특성

예측변수로는 Burt et al. (2001), Phinney et al. (2007), Frederick et al. (2014), Kang (2018) 등 다수의 연구를 통하여 주거불안정에 영향을 미치는 것으로 나타난 나이, 성별, 학력, 경제적 상황 등 가구특성과 주택의 유형, 점유 형태, 임대료, 주거만족도 등 주택특성을 설정하였다. 데이터 중심 접근법에서는 기존에 검증이 되지 않았던 변수의 영향을 탐색적으로 살펴볼 수 있다. 따라서 구득이 가능한 가구 및 주택특성을 추가로 포함하였다.

구체적으로 가구특성에서는 나이(*HAGE*), 성별(*HSEX*), 학력(*HEDU*), 소득(*HINC*), 자산(*HASS*), 부채(*HDEB*), 수급여부(*HREC*) 등을, 주택특성에서는 주택의 유형(*TYPE*), 위치(*POS*), 경과연수(*YEAR*), 방 개수(*ROOM*), 면적(*SIZE*), 임대료(*RENT*), 주택성능만족도(*PERF*), 주거환경만족도(*ENVI*), 임대 유형(*PUPR*), 점유형태(*TENU*) 등을 변수로 고려하였다. 이 중 일부는 모형의 과도한 복잡성을 제한하기 위하여 명목척도, 서열척도 등으로 변환하였다. 하지만 본 연구에서 활용하는 방법론은 변수의 측정 척도에 대하여 강건성(robustness)이 있으므로 통상적인 분석에서와 달리 소득, 자산, 부채, 면적, 임대료 등 규모의 변수를 로그로 변환하지 않았다. 한편 임대료의 경우, 우리나라에서는 전세, 보증부 월세, 무보증 월세 등 지불 방법이 다양하므로 보증금을 모두 월세로 계산한 환산임대료, 즉 환산한 월세로 측정하였다. 한국부동산원이 제공한 2019년 3분기 수도권 전월세전환율인 4.7%를 적용하여 환산하였다.

본 연구에서 활용한 결과변수와 예측변수에 대한 구체적인 설정 및 측정 등 개요는 <표 2>와 같다.

<표 2> 변수의 개요

구분	변수		측정	
결과변수	<i>ANXI</i>	주거불안	명목척도(0=‘불안하지 않다’, 1=‘불안하다’)	
예측변수	가구 특성	<i>HAGE</i>	나이	비율척도
		<i>HSEX</i>	성별	명목척도(0=여성, 1=남성)
		<i>HEDU</i>	학력	서열척도(1=초등학교 졸 이하; 5=대학원 졸 이상)
		<i>HINC</i>	소득	비율척도(월소득[만 원])
		<i>HASS</i>	자산	비율척도(총자산[만 원])
		<i>HDEB</i>	부채	비율척도(총부채[만 원])
		<i>HREC</i>	수급 여부	명목척도(0=미해당, 1=해상)
	주택 특성	<i>TYPE</i>	주택유형	서열척도(1=기타, 2=다세대·연립주택, 3=아파트)
		<i>POS</i>	위치	명목척도(0=[반]지하층·옥탑, 1=지상층)
		<i>YEAR</i>	경과연수	명목척도(0=20년 이하, 1=20년 초과)
		<i>ROOM</i>	방 개수	비율척도
		<i>SIZE</i>	면적	비율척도(전용면적[㎡])
		<i>RENT</i>	임대료	비율척도(환산한 월세[만 원])
		<i>PERF</i>	주택성능만족도	등간척도(1=‘매우 불만족’; 4=‘매우 만족’)
		<i>ENVI</i>	주거환경만족도	등간척도(1=‘매우 불만족’; 4=‘매우 만족’)
		<i>PUPR</i>	임대유형	명목척도(0=공공임대주택, 1=민간임대주택)
		<i>TENU</i>	점유형태	명목척도(0=전세 외, 1=전세)

3) 기술통계량

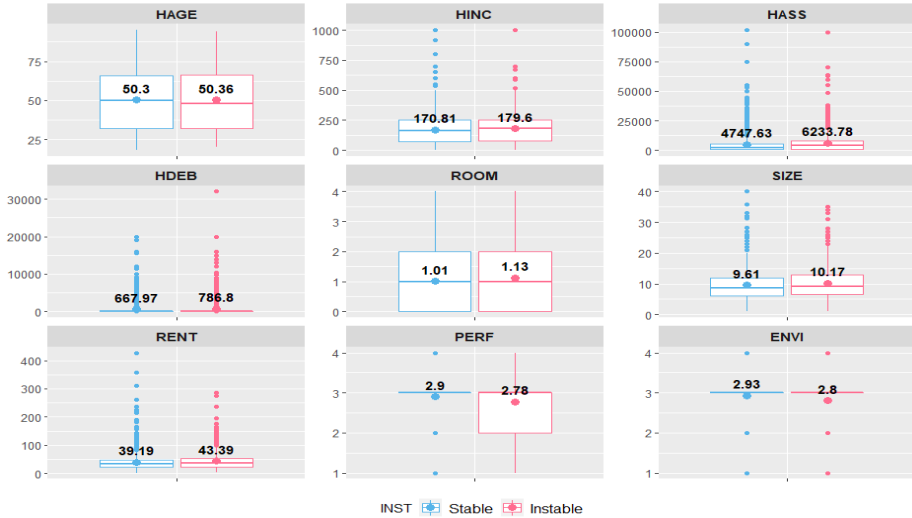
앞서 설정한 변수의 기술통계량을 통하여 수도권에 거주하는 1인 가구 임차인의 사회경제학적 특성을 확인할 수 있었다. 우선 비율척도, 등간척도로 측정한 예측 변수를 살펴본 결과(〈표 3〉), 평균 나이는 50.32세, 소득은 월 173만 원, 자산은 총 5,308만 원, 부채는 총 713만 원이었다. 또한 이들이 거주하는 주택의 방 개수는 1.06개, 면적은 9.82㎡, 임대료는 40만 원가량이었다. 주택성능만족도 (2.85점)와 주거환경만족도(2.88점)는 모두 2.5점 이상으로 불만족보다는 만족에 가까운 편으로 나타났다.

〈표 3〉 기술통계분석 결과

변수		<i>N</i>	<i>M</i>	<i>SD</i>	Min	Max
<i>HAGE</i>	나이	2,892	50.32	19.52	18	95
<i>HINC</i>	소득	2,892	174.13	109.08	0	1,000
<i>HASS</i>	자산	2,892	5,308.27	7,866.05	0	101,500
<i>HDEB</i>	부채	2,892	712.80	2,328.50	0	32,000
<i>ROOM</i>	방 개수	2,892	1.06	1.07	0	4
<i>SIZE</i>	면적	2,892	9.82	5.09	1	40
<i>RENT</i>	임대료	2,892	40.78	29.72	1.95	427.50
<i>PERF</i>	주택성능만족도	2,892	2.85	0.58	1	4
<i>ENVI</i>	주거환경만족도	2,892	2.88	0.53	1	4

이러한 특성이 주거불안을 경험하는 집단(이하 주거불안 집단, *Instable*)과 그렇지 않은 집단(이하 주거안정 집단, *Stable*) 간에 차이가 있는지 확인하였다 (〈그림 3〉). 또한 이러한 차이가 통계적으로 유의미한지 살펴보기 위하여 Welch의 독립표본 *t*-검정을 실시하였다. 분석 결과는 두 집단 사이에 자산, 방 개수, 면적, 임대료, 주택성능만족도, 주거환경만족도 등에 유의한 차이가 있음을 보여 주었다. 구체적으로 주거불안 집단이 주거안정 집단과 비교하여 총자산이 많고 ($t(2141.7)=4.86$, $p<.001$, $d=1.486.15$), 방 개수가 많고($t(2295.2)=2.89$, $p<.01$, $d=0.12$), 전용면적이 넓었다($t(2211.2)=2.81$, $p<.01$, $d=0.55$). 하지만 오히려 주거불안 집단의 임대료는 더 높았고($t(2252.1)=3.66$, $p<.001$, $d=4.19$), 주택성능 만족도($t(2173.1)=-5.55$, $p<.001$, $d=-0.13$)와 주거환경만족도($t(2092.5)=-6.12$, $p<.001$, $d=-0.13$)는 더 낮았다.

〈그림 3〉 주거불안 경험 여부에 따른 사회경제학적 특성의 차이



서열척도와 명목척도로 측정된 결과변수와 나머지 예측변수에 대하여 빈도분석을 실시한 결과는 〈표 4〉와 같다. 주거불안을 경험하고 있다는 응답은 1,091가구(37.72%), 그렇지 않다는 응답은 1,801가구(62.28%)로 나타났다. 주택유형은 다가구·단독주택 등을 포함한 기타의 응답이 1,802가구(62.31%)로 가장 많았다. 또한 절반에 가까운 1,296가구(44.81%)는 지어진 지 20년이 초과한 구옥(舊屋)에 거주하고 있었다. 주택의 임대유형은 대부분 민간임대주택(2,455가구, 84.89%)이었으며, 점유형태는 약 3분의 2가 월세·보증부월세 등(1,902가구, 65.77%)인 것으로 나타났다. 이외의 변수에서는 특이점을 발견할 수 없었다.

〈표 4〉 빈도분석 결과

변수	범주(값)	빈도	구성비(%)	누적구성비(%)
ANXI	주거불안			
	‘불안하지 않다’(=0)	1,801	62.28	62.28
	‘불안하다’(=1)	1,091	37.72	100.00
HSEX	성별			
	여성(=0)	1,376	47.58	47.58
	남성(=1)	1,516	52.42	100.00

변수		범주(값)	빈도	구성비(%)	누적구성비(%)
HEDU	학력	초등학교 졸 이하(=1)	481	16.63	16.63
		중학교 졸(=2)	400	13.83	30.46
		고등학교 졸(=3)	1,066	36.86	67.32
		대학교 졸(=4)	903	31.22	98.55
		대학원 졸 이상(=5)	42	1.45	100.00
HREC	수급 여부	미해당(=0)	2,413	83.44	83.44
		해당(=1)	479	16.56	100.00
TYPE	주택유형	기타(=1)	1,802	62.31	62.31
		다세대 · 연립주택(=2)	508	17.57	79.88
		아파트(=3)	582	20.12	100.00
POSI	위치	(반)지하층 · 옥탑(=0)	240	8.30	8.30
		지상층(=1)	2,652	91.70	100.00
YEAR	경과연수	20년 이하(=0)	1,596	55.19	55.19
		20년 초과(=1)	1,296	44.81	100.00
PUPR	임대유형	공공임대주택(=0)	437	15.11	15.11
		민간임대주택(=1)	2,455	84.89	100.00
TENU	점유형태	전세 외(=0)	1,902	65.77	65.77
		전세(=1)	990	34.23	100.00

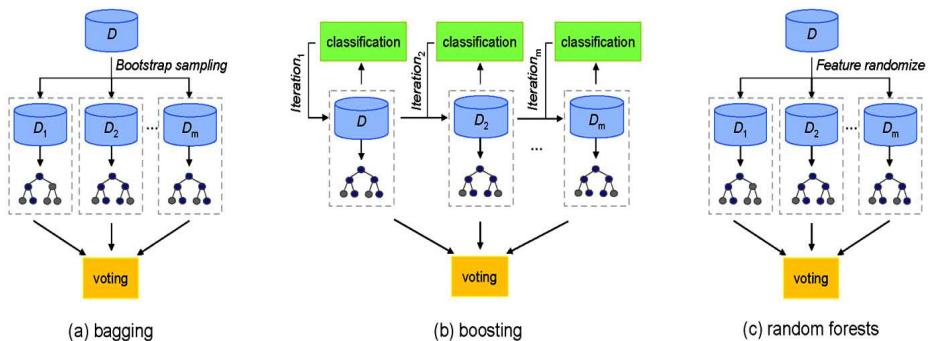
3. 모형

본 연구의 결과변수는 이항형이므로 분류모형(classification model)이 필요하다. 본 연구에서는 분류트리(classification tree)를 기반으로 앙상블 기법을 적용하였다. 앙상블 기법은 더 나은 예측성능을 얻기 위하여 다수의 분류기를 활용하는 알고리즘으로(Rokach, 2010) 배깅(bagging), 랜덤 포레스트, 부스팅(boosting)이 대표적이다. 이들은 대체로 단일 분류기보다 우수한 성능을 제공하는 것으로 알려져 있다(Kittler et al, 1998). 각 알고리즘에 대한 설명은 다음과 같다.

1) 분류트리

분류트리의 목적은 데이터를 가능한 작고 동질적인 집단으로 분할하는 것이다 (Breiman et al., 1984). 의사결정나무의 일종인 분류트리는 양적 결과변수를 가진 데이터에 활용하는 회귀트리(regression tree)와 달리 질적 결과변수의 예측에 활용한다. 구체적으로 분류트리는 각 표본이 속한 영역에서 훈련 표본이 가장 많이 포함된 클래스로 예측한다. 분류트리는 더미변수를 만들지 않고도 질적 예측변수를 쉽게 처리할 수 있고, 고차원의 상호작용효과를 감지할 수 있으며, 비선형적 상관관계를 식별하는 데 뛰어난 성능을 보인다는 장점이 있다(Sutton, 2005; James et al., 2013). 하지만 분류트리는 예측정확도가 낮는데, 이는 분류트리를 빌딩블록(building block)으로 활용하는 앙상블 기법을 활용하여 보완할 수 있다. 각 기법에 대한 개략도는 <그림 4>와 같다.

<그림 4> 앙상블 기법에 대한 개략도(Yang et al. (2010))



2) 배깅

앙상블 기법의 초기적 알고리즘인 배깅은 ‘부트스트랩 집계(bootstrap aggregating)’의 별칭으로 Breiman (1996)이 의사결정나무의 불안정성을 개선하기 위하여 제안하였다(Bühlmann, 2012). 배깅은 다음과 같은 장점이 있다. 첫째,

4) 해당 논문의 출판사인 Bentham Science 社의 사용 허가를 받았다. (This figure is used with permission from Bentham Science Publishers.)

의사결정나무는 높은 분산이 문제인데, 배깅은 다수의 트리를 집계하므로 예측 분산을 안정적으로 감소시킬 수 있다. 둘째, 배깅은 아웃 오브 백(out-of-bag, OOB) 표본을 활용하여 자체적인 오차추정값을 제공하므로 교차검증(cross validation, CV)이나 시험세트(training set)를 활용한 예측성능의 확인이 별도로 필요하지 않다(James et al., 2013; Kuhn & Johnson, 2013). 하지만 모든 앙상블 기법이 그러하듯 다수의 트리를 집계하므로 예측의 정확도는 향상되나 개별 변수나 모형에 대한 해석력은 낮아진다.

3) 랜덤 포레스트

랜덤 포레스트의 개념은 일찍부터 Amit & Geman (1997), Ho (1998) 등에 의해 제안되었으나 이를 알고리즘화한 것은 Breiman (2001)이었다. 랜덤 포레스트는 배깅의 알고리즘을 보완한 것인데, 핵심은 트리 간의 상관성을 제거했다는 점이다(Kuhn & Johnson, 2013). 배깅과 마찬가지로 부트스트래핑한 표본으로 만든 다수의 트리를 조합하지만, 더욱 우수한 예측성능을 도출하기 위하여 트리에 포함할 예측변수 역시 무작위로 추출한다. 각 트리에 포함할 예측변수의 개수로는 보통 전체 예측변수 개수의 제곱근 근삿값을 활용한다(Hastie et al., 2009).

4) 부스팅

부스팅 기법은 Schapire (1990), Freund (1995)에 의하여 탄생하였다. 배깅과 랜덤 포레스트는 다수의 트리를 집계하는 데에 반하여, 부스팅은 우선 하나의 트리를 만들고 해당 트리의 잔차 정보를 바탕으로 순차적으로 모형을 수정하고 개선한다. 따라서 부스팅은 약 분류기(weak classifier)를 강 분류기(strong classifier)로 만드는 데 최적화된 기법이라 할 수 있다(James et al., 2013). 알고리즘으로는 Freund & Schapire (1996)의 에이다부스트(AdaBoost)와 Friedman et al. (2000)의 경사부스팅머신(gradient boosting machine, GBM) 등이 있는데, 본 연구에서는 가중치 부여 메커니즘의 차이에 따라 조금 더 성능이 우수한 것으로 알려진 GBM을 활용하였다.

IV. 모델링

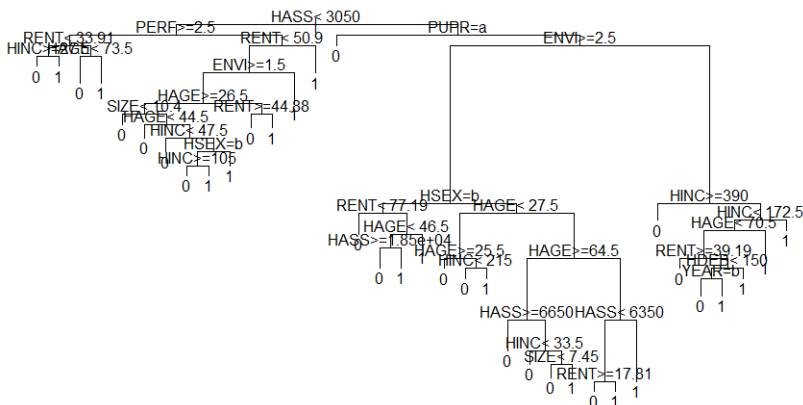
1. 모형구축

모델링에 앞서 데이터셋을 훈련세트(training set), 시험세트로 분할하였다. 훈련세트는 모형을 구축하는 데, 시험세트는 모형을 평가하고 선택하는 데 이용한다. 통상적인 관례에 알맞게 9:1의 비율로 나누었다. 데이터셋의 크기가 충분히 크지 않아 검증세트(validation set)는 따로 만들지 않았다.⁵⁾ 이 데이터셋을 활용하여 앞서 소개한 네 개의 알고리즘으로 네 개의 모형을 구축하였다. 이후 예측성능을 비교평가하여 가장 예측력이 강력한 모형을 선택한다. 모델링에는 오픈소스 R을 활용하였다.

1) 분류트리

CART(classification and regression trees) 알고리즘을 활용하는 ‘rpart’ 패키지 4.1-15 버전을 이용하여 분류트리 모형을 구축하였다. 최초 구축한 빅 트리(big tree)의 단말노드(terminal node, leaf node)는 36개, 분류 정확도는 61.94%로 나타났다(<그림 5>).

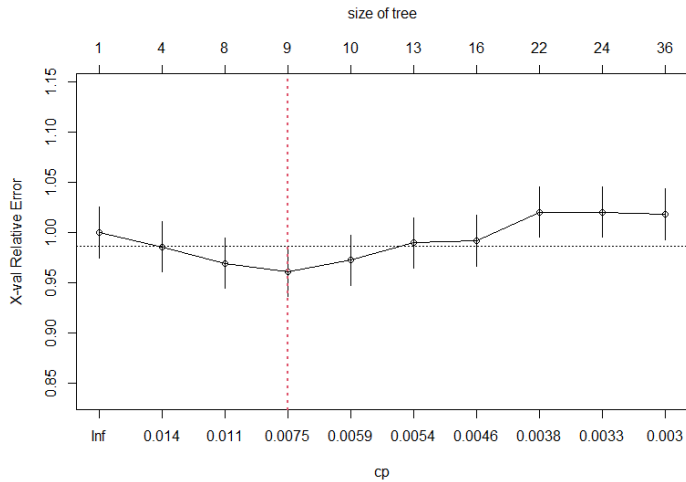
<그림 5> 최초 구축한 빅 트리



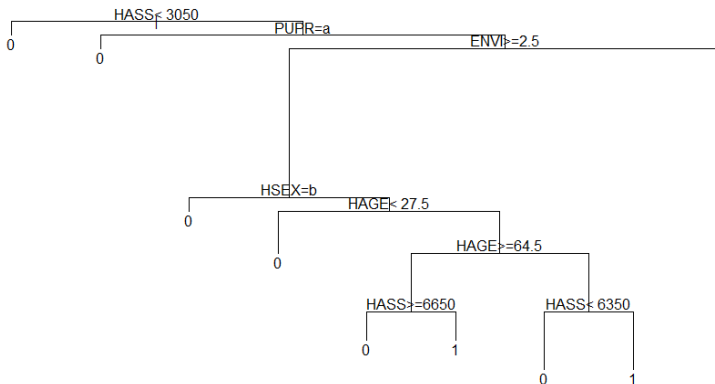
5) 분류트리는 가지치기(pruning) 과정에서 자체적인 검증을 수행하고, 배깅과 랜덤 포레스트는 별도의 OOB 표본으로 검증한 오류를 제공하며, 부스팅은 모형구축 과정에서 교차검증이 가능하기 때문이기도 하다.

빅 트리는 탐욕 알고리즘(greedy algorithm)에 따라 최적화된 모형으로 훈련 세트에 과적합(over-fitting)할 가능성이 크므로(Kuhn & Johnson, 2013) 가지 치기가 필요하다. Breiman et al. (1984)이 제안한 바에 따라 복잡성 파라미터 (complexity parameter, CP)를 활용한 가지치기를 실시하였다. 복잡성 파라미터의 값을 증가시키며, 즉 트리의 크기를 감소시키며 오류율이 최소화되는 지점을 확인한 결과, 단말노드가 9개인 하위트리(subtree)가 가장 적합한 것으로 나타났다(<그림 6>). 이에 따라 <그림 7>과 같은 트리를 구축하였다. 분류 정확도는 64.01%로 빅 트리과 비교하여 예측성능이 소폭 향상되었음을 확인할 수 있었다.

<그림 6> 복잡성 파라미터를 활용한 가지치기



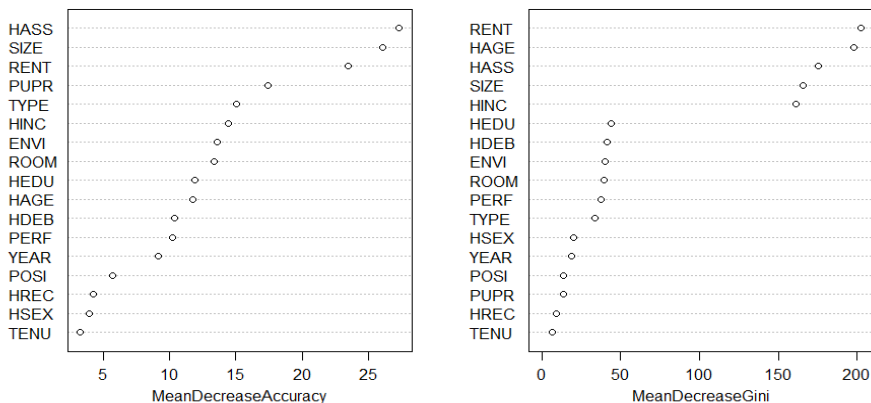
<그림 7> 가지치기를 통하여 도출한 트리



2) 배깅

배깅 모형의 구축에는 ‘randomForest’ 패키지 4.6-14 버전⁶⁾을 이용하였다. 트리의 개수는 500개⁷⁾로 설정하였다. 모델링 결과, 분류 정확도는 69.55%로 나타났다. 이는 분류트리의 64.01%보다 약 5.5%p 높은 것으로 배깅의 예측성능이 더 우수하다는 것을 의미한다. 한편 앙상블 기법에서는 각 예측변수가 분류에 얼마나 기여하는지 확인할 수 있다. 이는 각 변수로 인한 분할이 분류의 정확도나 지니지수(Gini index)를 얼마나 감소시키는가의 정도로 계산한다. 정확도평균 감소(mean decrease accuracy)를 기준으로 살펴본 결과, 총자산, 면적, 임대료, 임대유형, 주택유형 등이 주거불안을 예측하는 데 가장 중요한 상위 5개의 변수인 것으로 나타났다(<그림 8>).

<그림 8> 배깅 모형에 이용한 예측변수의 정확도평균감소와 지니지수평균감소



3) 랜덤 포레스트

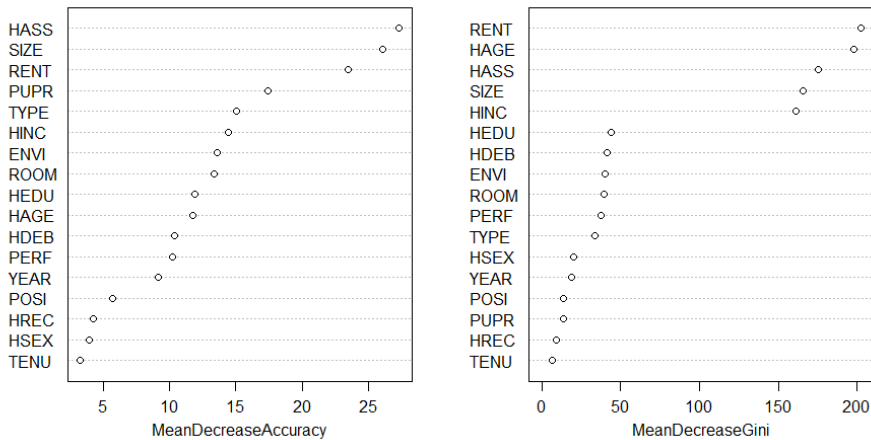
랜덤 포레스트 모형의 구축에는 배깅과 마찬가지로 ‘randomForest’ 패키지 4.6-14 버전을 이용하였다. 각 트리에 전체 예측변수 개수 17개의 제곱근 근삿값인 4개의 예측변수를 투입하였고 트리의 개수는 배깅과 마찬가지로 500개로 설정하였다. 모델링을 실시한 결과, 분류 정확도는 72.66%로 나타나 배깅 모형

6) 파라미터를 조절하여 랜덤 포레스트와 배깅 알고리즘을 모두 구현할 수 있다.

7) 트리의 개수가 50~500개일 때 최적의 성능을 내는 것으로 알려져 있다(Boehmke & Greenwell, 2019).

보다 예측성능이 더 개선되었음을 확인할 수 있었다. 또한 정확도평균감소를 기준으로 총자산, 면적, 주택유형, 임대료, 임대유형이 주거불안을 예측하는 데에 가장 중요한 5개의 변수인 것으로 나타났다(<그림9>).

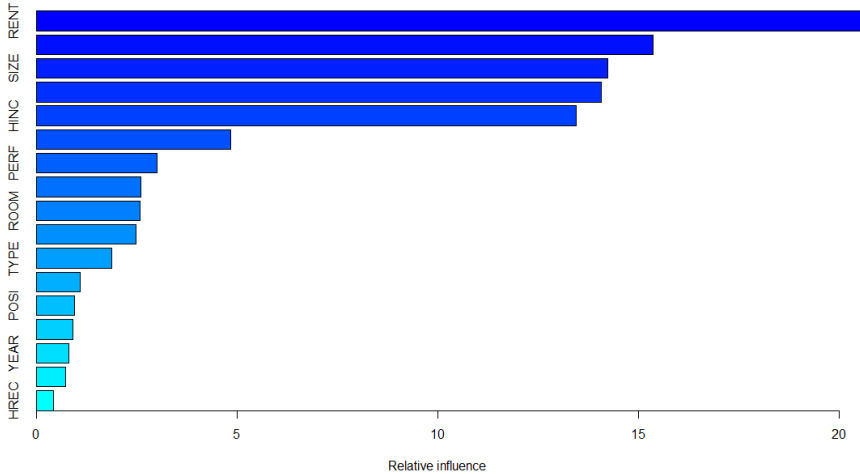
〈그림 9〉 랜덤 포레스트 모형에 이용한 예측변수의 정확도평균감소와 지니지수평균감소



4) 부스팅

부스팅 모형의 구축에는 경사부스팅머신 알고리즘을 지원하는 ‘gbm’ 패키지 2.1.8 버전을 이용하였다. 트리의 개수는 배깅, 랜덤 포레스트와 마찬가지로 500개로 설정하였으며, 초모수(hyperparameter)인 수축파라미터(shrinkage parameter)는 0.3, 상호작용 깊이(interaction depth)는 2로 튜닝(tuning)하였다. 또한 교차 검증 파라미터에는 10의 값을 부여하여 10겹 교차검증(10-fold CV)을 하도록 하였다. 모델링 결과, 분류 정확도는 67.82%로 기본적인 분류트리 모형보다는 높았으나 배깅, 랜덤 포레스트 모형보다는 낮은 것으로 나타났다. 대체로 부스팅이 랜덤 포레스트보다 더 나은 성능을 제공하지만, 과적합에 민감하다(Vezhnevets & Barinova, 2007)는 이유 때문일 수 있다. 한편 상대적인 영향력(relative influence)을 기준으로 임대료, 총자산, 면적, 나이, 월소득이 주거불안을 예측하는 데에 가장 중요한 5개의 변수인 것으로 나타났다(<그림 10>). 아울러 위치, 임대유형, 경과 연수, 점유형태 등의 변수는 상대적인 영향력이 0에 가까운 것으로 나타났다.

〈그림 10〉 부스팅 모형에 사용된 예측변수의 상대적인 영향력



3. 모형평가 및 선택

모형의 예측성능 평가의 지표로는 분류 예측이 정확하게 되었는지를 측정하는 정확도와 PPV(positive predictive value)를 표현하는 정밀도(precision) 그리고 의사결정 이론(decision theory)의 측면에서 활용하는 ROC(receiver operating characteristics) 곡선 아래 면적, 즉 AUC(area under the ROC curve)를 함께 활용하여 강건한 결과를 도출하고자 하였다.

여기서 정확도(식 (1))는 전체 케이스에서 주거불안의 경험을 정확하게 분류한 케이스의 비율이다. 정밀도(식 (2))는 주거불안을 경험할 것으로 예측된 케이스 중 실제 주거불안을 경험하고 있는 케이스의 비율이다. 따라서 두 지표 모두 높으면 높을수록 예측성능이 우수한 것이다.

$$Classification\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad \cdots (1)$$

$$Precision = \frac{TP}{TP + FP} \quad \cdots (2)$$

ROC 그래프의 가로축은 1-특이도(specificity), 세로축은 민감도(sensitivity)를 표현하므로 AUC는 1에 가까울수록 예측성능이 우수한 것으로 판단할 수 있다.

만약 AUC가 1이면 특이도와 민감도가 모두 1이므로 완벽한 예측모형이다. 이에 반해 0.5에 가까울수록 무작위의 예측과 차별성을 갖지 못하는 불량한 모형이다.

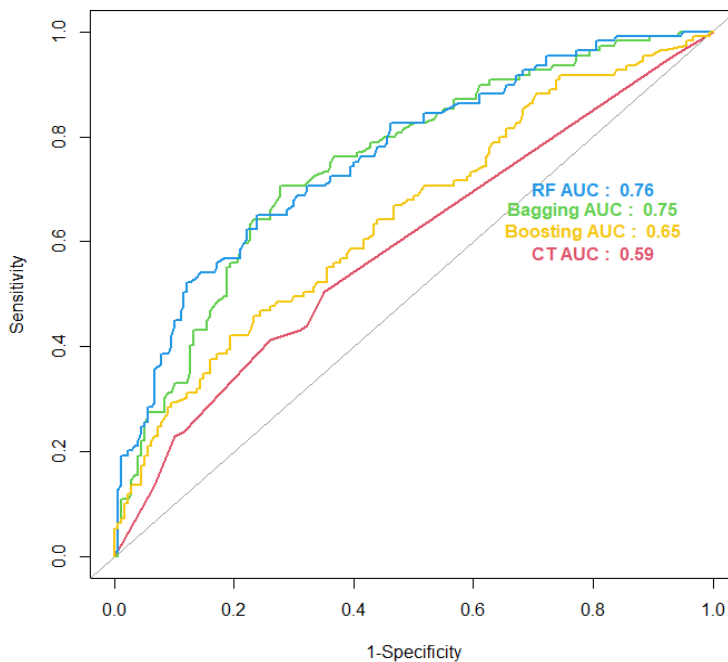
정확도, 정밀도, AUC 등 세 개의 지표를 계산한 결과는 <표 5>와 같다. AUC 그래프는 <그림 11>에 표현하였다. 이를 종합하면 모든 지표에서 랜덤 포레스트 모형의 예측능력이 가장 우수한 것으로 나타난 것으로 볼 수 있다. 따라서 최종적으로 랜덤 포레스트로 구축한 모형을 수도권 1인 가구 임차인의 주거불안 예측 모형으로 선택하였다.

<표 5> 모형평가 결과

모형	정확도	정밀도	AUC
분류트리	64.01%	55.32%	0.59
배깅	69.55%	63.64%	0.75
랜덤 포레스트	72.66%	72.06%	0.76
부스팅	67.82%	59.09%	0.65

주: AUC=area under the ROC curve

<그림 11> 네 모형의 AUC 그래프



V. 결 론

본 연구에서는 앙상블 기반의 머신러닝을 활용하여 수도권 1인 가구 임차인의 주거불안 예측모형을 개발하였다. 저자는 수도권 1인 가구의 주거불안을 스크리닝할 수 있는 예측 모델링(predictive modeling)이 필요하다고 보았기 때문이다. 데이터세트는 『2019년도 주거실태조사』의 응답을 전처리하여 구축하였고, 모델링에는 분류트리, 배깅, 랜덤 포레스트, 부스팅 등의 알고리즘을 적용하였다.

수도권 1인 가구 임차인의 주거불안 예측모형을 개발한 결과, 랜덤 포레스트로 구축한 모형이 가장 우수한 성능을 보이는 것으로 나타났다. 이 모형은 분류 정확도가 72.66%, 정밀도가 72.06%, AUC가 0.76으로, 본 연구가 주거불안 예측 모형의 개발에 관한 초기적 시도임을 감안하였을 때 양호한 수준의 성능을 확보한 것으로 평가할 수 있다. 서울·인천·경기 등 수도권 지자체에서 이를 실제 정책과정에 활용한다면 1인 가구 임차인이 주거위기의 상황에 빠지기 전 예방적 조치를 취할 수 있을 것이다. 또한 주거불안정을 측정할 수 있는 객관적인 지표와 결합하여 활용한다면 주거안전망 체계의 일부로도 활용할 수 있다. 이는 최신의 방법론을 활용하여 정책 수요를 선제적으로 파악하고 대응하려는 접근이라는 점에서 정부가 장려하는 적극 행정의 취지에 부합한다.

앙상블 기법을 활용한 본 연구에서는 어떠한 특성이 주거불안 경험의 분류에 더 크게 기여하는지 식별할 수 있었다. 구체적으로 배깅 모형에서는 총자산, 면적, 임대료, 임대유형, 주택유형 등, 랜덤 포레스트 모형에서는 총자산, 면적, 주택유형, 임대료, 임대유형 등, 부스팅 모형에서는 임대료, 총자산, 면적, 나이, 월소득 등이 각각 주거불안 경험의 예측에 가장 중요한 상위 5개 변수인 것으로 나타났다. 이 중 특히 총자산, 임대료, 면적 등은 일관되게 중요한 변수인 것으로 드러났다. 이는 앞선 t-검정의 결과와 함께 살펴보면 흥미로운 해석이 가능하다. 주거불안 집단이 주거안정 집단과 비교하여 총자산은 많았던 데 비하여, 임대료가 높고, 면적이 넓은 주택에 거주하고 있었기 때문이다. 이들 세 변수는 주거불안 경험의 분류에 크게 기여한 데 반하여, 평균의 차이에서는 통념적 사고와 다른 경향을 나타냈다. 거칠게 해석하면 수도권 1인 가구 임차인이 주관적으로 경험하고 느끼는 주거불안은 객관적인 주거불안정의 지표와 괴리될 수 있다는 뜻이다. 개인의 경제적 수준에 따라 주거의 기대치가 다르다는 사실을 견지한다면 직

관적으로 받아들이기 어렵지 않다. 이는 주거불안정의 개념이 주거환경에 대한 통제권을 쥐지 못하는 광범위한 상황이라는 개념적 정의(Kang, 2019)와 궤를 함께한다. 따라서 이들을 위한 주거복지정책은 현행대로 최저주거기준, 주거비 등 객관적인 지표에 기반하되, 주관적 경험에 대한 보고는 보조적 지표로 활용하는 것이 바람직하다. 아울러 객관적인 상황과 괴리된 판단으로 불안을 겪는 가구를 위하여 주거 관련 상담서비스를 제공할 필요가 있다. 최근 확대되는 지자체의 1인 가구 지원센터 등과 연계를 도모할 수 있다.

본 연구는 수도권 1인 가구 임차인의 주거문제에 대한 예방적 조치가 필요한 현시점에 수행되었다는 점에서 의의가 있다. 하지만 소위 ‘블랙박스(black box)’라 불리는 알고리즘을 가진 머신러닝을 활용하였기에, 인과적 접근에서와 같이 변수의 특성에 대한 풍부한 해석은 불가능하였다. 물론 저자가 앞서 밝혔듯 두 접근의 목표가 다르므로 이를 연구의 결점이라 하기는 어렵다. 하지만 로지스틱 회귀모형 등 전통적인 기법을 함께 활용하고 이를 비교한다면 또 다른 시사점을 도출할 수 있을 것이다.

참고문헌

1. 강은나·이민홍(2016), “우리나라 세대별 1인가구 현황과 정책과제,” 『보건복지포럼』, (234): 47-56.
2. 구한민·김갑성. (2021). “공공임대주택 거주자의 주거만족도 형성과정: 특성요인에 대한 만족도의 매개효과,” 『주택연구』, 29(1): 33-78.
3. 구한민·김갑성. (2023). “공공임대주택, 청년 주거의 디딤돌인가?: 임대주택 유형이 자가보유의식에 미치는 영향과 주거만족도의 매개효과,” 『주택연구』, 31(2): 43-71.
4. 구한민·홍사흠. (2023). “지역일자리, 대체될 것인가 지속할 것인가?: 지역특성이 기술진보에 따른 일자리 대체 가능성에 미치는 영향,” 『대한지리학회지』, 58(2): 135-146.
5. 권연화·최열(2018), “1인가구의 연령별 특성에 따른 주거상·하향 이동에 관한 연구,” 『국토연구』, 99: 97-112.

6. 김옥연·문영기(2009), “1인가구 주거실태 분석: 서울시 1인가구를 중심으로,” 『주거환경』, 7(2): 37-53.
7. 김윤영·김아진(2017), 『인천시 1인가구 특성 및 정책수요 연구』, 인천연구원.
8. 김은미·김상봉·조은서(2020), “기계학습을 활용한 주택매도 결정요인 분석 및 예측모델 구축,” 『지적과 국토정보』, 50(1): 181-200.
9. 김정기·구한민·김갑성(2023), “공유주거의 고급화와 차별화: 직장인 청년 1인가구의 셰어하우스 수요 특성 탐색,” 『도시재생』, 9(2): 24-42.
10. 김재익(2013), “1인가구의 생애주기별 이질성과 공간적 분포특성,” 『주택연구』, 21(3): 63-81.
11. 나성호·김종우(2019), “공공데이터를 활용한 아파트 매매 가격 결정 모형의 예측능력 비교: 서울 강남구 지역을 중심으로,” 『한국지적정보학회지』, 21(1): 3-12.
12. 박미선·이재춘·김근용·이운상·천현숙(2017), 『1인가구 증가에 따른 주택정책 대응방안 연구』, 국토연구원.
13. 반정호(2012), “1인 가구의 사회·경제적 특성과 변화,” 『노동리뷰 2012년 4월호』, 55-67.
14. 배성완·유정석(2018), “기계 학습을 이용한 공동주택가격 추정: 서울 강남구를 사례로,” 『부동산학연구』, 24(1): 69-85.
15. 변미리·최정원·박민진·이혁준·김진아(2015), 『서울특별시 1인가구 대책 정책연구』, 서울연구원 미래연구센터.
16. 손철(2011), “수도권의 공간적 주택하위시장 식별에 대한 연구,” 『국토연구』, 70: 151-166.
17. 송호창·김태호·이주형(2008), “주상복합아파트의 주택규모별 가격결정요인 분석,” 『서울도시연구』, 9(3): 79-92.
18. 신상영(2010), “1인가구 주거지의 공간적 분포에 관한 연구: 서울시를 사례로,” 『국토계획』, 45(4): 81-95.
19. 이동훈(2012), 『1인가구를 위한 소형임대주택 공급확대 방안』, 서울연구원.
20. 이석준(2019), 『데이터마이닝을 통한 주택 하위시장 구분 및 주택가격 예측』, 서울대학교 환경대학원 박사학위논문.

21. 이준용·최미화·이상엽(2007), “데이터 마이닝 적용을 통한 아파트 가격 예측에 관한 연구,” 『국토계획』, 42(4): 135-148.
22. 정기성(2021), “신경망 분석을 활용한 비주택 거주 청년 1인 가구의 공유주택 거주 의사에 미치는 영향 요인 분석,” 『주택연구』, 29(2): 37-58.
23. 정인·오상엽(2020), 『2020 한국 1인가구 보고서』, KB금융 경영연구소 1인가구 연구센터.
24. 정진호·김중협·추재호·이승훈·현창택(2017), “랜덤 포레스트 기반 공동주택 공용관리비 추산모델,” 『대한건축학회 논문집-계획계』, 33(3): 19-27.
25. 최석현·김재신(2017), 『경기도민 삶의 질 조사 VII: 1인가구』, 경기연구원.
26. 통계청(2022a), 『2021년 인구주택총조사 결과: 등록센서스 방식(보도자료)』, 2022.07.28.
27. 통계청(2022b), 『2022 통계로 보는 1인가구(보도자료)』, 2022.12.07.
28. 통계청(2022c), 『수도권 1인 가구 특성 변화 및 현황 분석: 지난 10년간 1인 가구 특성 변화(보도자료)』, 2022.10.12.
29. 통계청(2022d), 『장래가구추계: 2020~2050년(보도자료)』, 2022.06.28.
30. 홍정의(2021), “랜덤 포레스트 알고리즘을 통한 주택 대량평가모형 연구,” 『부동산분석』, 7(1): 1-28.
31. Belcher, J. C.(1967), “The One-Person Household: A Consequence of the Isolated Nuclear Family?,” *Journal of Marriage and the Family*, 29(3): 534-540.
32. Boehmke, B., and Greenwell, B. M.(2019). *Hands-On Machine Learning with R*, CRC Press.
33. Breiman, L.(1996), “Bagging Predictors,” *Machine learning*, 24(2): 123-140.
34. Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A.(1984), *Classification and Regression Trees*. CRC press.
35. Bühlmann, P.(2012), Bagging, Boosting and Ensemble Methods. In Gentle, J. E., Härdle, W. K., and Mori, Y.(Eds), *Handbook of Computational Statistics (pp. 985-1022)*, Springer.

36. Burt, M. R., Aron, L. Y., and Lee, E.(2001), *Helping America's Homeless: Emergency Shelter or Affordable Housing?*, The Urban Institute Press.
37. Fan, G. Z., Ong, S. E., and Koh, H. C.(2006), "Determinants of House Price: A Decision Tree Approach," *Urban Studies*, 43(12): 2301-2315.
38. Frederick, T. J., Chwalek, M., Hughes, J., Karabanow, J., and Kidd, S.(2014), "How Stable is Stable? Defining and Measuring Housing Stability," *Journal of Community Psychology*, 42(8): 964-979.
39. Freund, Y.(1995), "Boosting a Weak Learning Algorithm by Majority," *Information and Computation*, 121(2): 256-285.
40. Freund, Y., and Schapire, R. E.(1996), "Experiments with a New Boosting Algorithm," *Proceedings of the Thirteenth International Conference*, 1-9.
41. Friedman, J., Hastie, T., and Tibshirani, R.(2000), "Additive Logistic Regression: A Statistical View of Boosting," *The Annals of Statistics*, 28(2): 337-407.
42. Gupta, R., Marfatia, H. A., Pierdzioch, C., and Salisu, A. A.(2021), "Machine Learning Predictions of Housing Market Synchronization Across US States: The Role of Uncertainty," *The Journal of Real Estate Finance and Economics*, 64(6): 523-45.
43. Hastie, T., Tibshirani, R., and Friedman, J.(2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Science & Business Media.
44. James, G., Witten, D., Hastie, T., and Tibshirani, R.(2013), *An Introduction to Statistical Learning*, Springer.
45. Kang, S.(2019), "Why Low-Income Households Become Unstably Housed: Evidence from the Panel Study of Income Dynamics," *Housing Policy Debate*, 29(4): 559-587.
46. Kittler, J., Hatef, M., Duin, R. P., and Matas, J.(1998), "On Combining Classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3): 226-239.

47. Kleit, R. G., Kang, S., and Scally, C. P.(2016), "Why Do Housing Mobility Programs Fail in Moving Households to Better Neighborhoods?," *Housing Policy Debate*, 26(1): 188-209.
48. Kobrin, F. E.(1976), "The Fall in Household Size and the Rise of the Primary Individual in the United States," *Demography*, 13(1): 127-138.
49. Kuhn, M., and Johnson, K.(2013), *Applied Predictive Modeling*, Springer.
50. O'Campo, P., Daoud, N., Hamilton-Wright, S., and Dunn, J.(2016), "Conceptualizing Housing Instability: Experiences with Material and Psychological Instability among Women Living with Partner Violence," *Housing Studies*, 31(1): 1-19.
51. Özsoy, O., and Şahin, H.(2009), "Housing Price Determinants in Istanbul, Turkey: An Application of the Classification and Regression Tree Model," *International Journal of Housing Markets and Analysis*, 2(2): 167-178.
52. Pavao, J., Alvarez, J., Baumrind, N., Induni, M., and Kimerling, R.(2007), "Intimate Partner Violence and Housing Instability," *American Journal of Preventive Medicine*, 32(2): 143-146.
53. Phinney, R., Danziger, S., Pollack, H. A., and Seefeldt, K.(2007), "Housing Instability among Current and Former Welfare Recipients," *American Journal of Public Health*, 97(5): 832-837.
54. Rokach, L.(2010), "Ensemble-Based Classifiers," *Artificial Intelligence Review*, 33: 1-39.
55. Rollins, C., Glass, N. E., Perrin, N. A., Billhardt, K. A., Clough, A., Barnes, J., Hanson, G. C., Barnes, J., and Bloom, T. L.(2012), "Housing Instability is as Strong a Predictor of Poor Health Outcomes as Level of Danger in an Abusive Relationship: Findings from the SHARE Study," *Journal of Interpersonal Violence*, 27(4): 623-643.
56. Schapire, R. E.(1990), "The Strength of Weak Learnability," *Machine Learning*, 5(2): 197-227.

57. Sutton, C. D.(2005), "Classification and Regression Trees, Bagging, and Boosting," *Handbook of Statistics*, 24, 303-329.
58. Vezhnevets, A., and Barinova, O.(2007), "Avoiding Boosting Overfitting by Removing Confusing Samples," *Proceedings of European Conference on Machine Learning 2007*, 430-441.
59. Yang, P., Hwa Yang, Y., B Zhou, B., and Y Zomaya, A.(2010), "A Review of Ensemble Methods in Bioinformatics," *Current Bioinformatics*, 5(4): 296-308.
60. Yi, C., and Kim, K.(2018), "A Machine Learning Approach to the Residential Relocation Distance of Households in the Seoul Metropolitan Region," *Sustainability*, 10(9): 2996.
61. [http://www.reb.or.kr/\(2023.04.27.\)](http://www.reb.or.kr/(2023.04.27.))