pISSN : 2465-9754 eISSN : 2508-1292 https://doi.org/10.30902/jrea.2021.7.1.1

*Journal of Real Estate Analysis*

*http://www.ejrea.org* *April 2021, Vol.7, No.1, pp. 1~28*

# 랜덤 포레스트 알고리즘을 통한 주택 대량평가모형 연구

# A Mass Appraisal Model on Residential Property with Random Forest Algorithm

홍정의*
Jengei Hong

## Abstract

This paper discusses how to apply the random forest algorithm in building a mass appraisal system for residential property and analyzes issues related to modelling process. The paper investigates the relationship between the random forest model and the complexities of housing market. Based on the findings, various qualitative analyses has been attempted for effective model design. The findings are summarized as followed:-. First, the random forest model is performative in capturing the non-linearity from sub-market and locational effects. The random forest model has significantly low average percentage error (approx. 4%) compared to with a linear Hedonic model (approx. 11%). Second, the random forest model can efficiently capture the locational effects only with locational information (coordinates), without proxy variables. Third, using dummy variables may reduce explanatory power of the model, compared to label indexes because the number of variables included in an operation increases. Fourth, the advantage from model complexity seems overwhelm the disadvantage from overfitting. Fifth, modellers are not required to ensure consistency in the time periods contained in a dataset.

**Keywords:** Machine learning, Random forest, Mass appraisal, Housing sub-markets, Locational effect

* 한동대학교 경영경제학부 조교수 | Assistant Professor, Management and Economics, Handong Global University | hwgh024@handong.edu |

## Ⅰ. 서론

주택은 주식 등의 금융자산과 달리 거래빈도가 낮고 각각 고유한 입지에 기반했다는 데에서 오는 이질성이 높다. 이는 주택담보대출에 기반한 금융상품의 설계나 주택가격지수 계산, 부동산 개발사업 추진, 적절한 세금의 산정 등 다양한 의사결정 과정에서 주택 가치의 잦은 감정평가가 불가피하다는 것을 의미한다. 크게 보면 두 가지의 감정평가 방식이 존재할 수 있는데, 하나는 전문성을 갖춘 감정평가사가 개별 주택의 적정 시장 가치를 직접 평가하는 것이고, 두 번째는 주택의 시장 가격을 평가하는 수리적-기술적 모형을 통해 시장에서 매겨졌을 가격을 추정하는 방식이다. 전통적으로 주택의 감정평가란 감정평가사를 통한 직접 추정을 의미하며, 이는 일반적으로 두 번째에 비해 정교하지만 불가피하게 비용과 시간이 많이 소요되는 방식이다. 그러므로 대량의 주택에 대한 가치를 빈번하게 추정해야 하는 상황에서는 이른바 주택가격 대량평가모형(mass-appraisal model)을 통한 가치 추정이 행해질 수 있다(Wang and Li, 2019; Zhou and Zhang, 2018). 예를 들어, 주택가격지수의 추정 시 동일한 주택의 시장 가격을 매시점 추적해야 한다. 이때, 모든 주택이 매시점 거래되는 것이 아니므로 거래되지 않은 주택의 시장 가격을 추정해야 하는 문제가 생긴다. 이는 일반적으로 대규모의 표본에 대한 예측이 필요하므로, 모형을 통한 가치 추정을 하게 된다. 이와 같은 대량 감정평가의 필요성은 주택관리/주택금융 시스템이 발전하고, 더 많은 자산에 대한 빈번한 가치 평가가 요구될수록 높아진다.

전통적으로 이러한 대량 평가모형은 선형 헤도닉 모형에 기반하여 이루어져 왔다. Rosen(1974)에 의해 최초로 소개된 헤도닉 모형은 특정한 조건 하에서 주택 가격을 형성하는 각 요소의 시장 가치를 개별적으로 추정할 수 있다는 이론적 기반을 제공하였다(Sheppard, 1999). 헤도닉 모형의 강점은 결과의 해석이 용이하고 직관적이며, 기술적으로 이미 잘 알려져 있는 선형다중회귀모형의 적용만으로도 주택 가격의 평가를 시도할 수 있다는 것이다. 한편, 이러한 강점은 동시에 약점이 되기도 한다. 전통적인 헤도닉 모형에서 가정하는 조건들이 시장에서 만족이 되지 않거나, 특히 변수간 관계에 내생성이 있거나 비선형성이 강한 경우 모형의 정확성이 저하될 수 있기 때문이다(Ramsey, 1969). 예를 들어, 하부 시장이 존재하거나 완전 경쟁 가정이 성립하지 않는 경우(Malpezzi, 2002), 지역 노동시장의 성격이나 이웃효과, 문화적 요인 등 지가에 영향을 미치나, 관찰 가능한 변수의 형태로 모형에 반영되지 않는 경우(Nesheim, 2002; Osland and Thorsen, 2008; Wheaton and Lewis, 2002; Zukin, 1987), 헤도닉 모형의 설명력은 훼손될 수 있다.

최근 4차 산업혁명기로 접어들면서 많은 데이터 과학 분석 기법이 소개되었고, 이러한 기법을 대량 주택 평가 모형의 설계에 적용하는 연구들도 급격히 성장하고 있다. 이처럼 데이터 과학 기술을 부동산 산업 및 연구 분야에 접목한 프롭테크(proptech)는 당분간 흐름을 이어갈 것으로 예상된다.

그 중 최근 가장 각광 받는 기계 학습 기법 중 하

 https://doi.org/10.30902/jrea.2021.7.1.1

나는 랜덤 포레스트 알고리즘이다. 랜덤 포레스트는 전통적인 데이터분석/예측 기법 중 하나인 의사결정나무 알고리즘의 앙상블 기법으로, 의사결정나무 알고리즘의 단점인 과적합 가능성을 극복하고 높고 안정적인 예측력을 갖는 것으로 알려져 있으며, 이에 기반한 부동산 대량평가모형이 헤도닉 모형에 비해 높은 정확성을 갖는다는 것이 여러 문헌에서 밝혀졌다(예를 들어, Antipov and Pokryshevskaya, 2012; Hong et al., 2020 등). 다만 랜덤 포레스트 기법을 활용한 분석은 아직 초기 단계라고 볼 수 있다. 기존 연구의 대부분은 랜덤 포레스트와 다른 모형의 예측력 비교에 초점을 맞추고 있는 반면, 랜덤 포레스트 모형을 어떻게 효율적으로 구현해야 하는지에 대해서는 지엽적으로만 다루고 있다.

본문의 주요 논점은 주택시장분석을 위해 랜덤 포레스트를 사용할 때 그 모형을 어떻게 효과적으로 구현할 수 있는지를 알아보는 것이다. 예를 들어 헤도닉 모형의 경우에도 함수형태(변수 관계를 자연로그로 표현하거나, 보다 복잡하게는 Box-cox 함수 등으로 표현하는 등), 변수의 표현방식(예를 들어 입지효과 표현을 위해 주요 시설과의 유클리드 거리를 직접 표현할 수도 있고, 행정구역/지역 소득 등을 더미화 하거나 공간 자기상관계수 등 포함할 수도 있음) 등 모형의 세부적 구현에 대한 이슈가 광범위하게 존재한다. 반면, 랜덤 포레스트와 같이 최근에 조명받기 시작한 방법론은 주택시장 분석에 있어 어떤 방식으로 모형을 구현하는 것이 효율적인지는 아직까지 알려진 바가 거의 없다. 랜덤 포레스트 역시 연구자가 설정한 모형의 세부사항에 따라 모형의 예측력과 활용성에 큰 차이가 나타날 수 있기 때문에, 모형의 활용이 보다 일반화되기 위해서는 이에 대한 논의가 반드시 필요할 것으로 보인다.

따라서 본문은 먼저 랜덤 포레스트 기법과 주택 가격 결정구조의 특징에 대해 고찰하였다. 헤도닉 모형과 달리, 랜덤 포레스트는 변수 간 관계를 묘사하는 특정한 함수가 없는 비직관적 구조를 가지고 있다. 그러므로 랜덤 포레스트가 변수(본문에서는 주택 가격)를 포착하는 방식을 먼저 이해할 수 있어야, 그러한 방식이 주택 시장이 가진 어떤 특성을 포착하는 데에 활용되는지를 논의할 수 있다.[1] 그리고, 이와 같은 랜덤 포레스트가 변수를 포착하는 방식과 주택 가격 결정 구조의 특성 간 관계에 기반하여, 연구자가 어떻게 효율적인 모형을 설정(specification)할 수 있는지를 논의할 수 있다.

그 다음으로, 본문은 위 논의된 내용을 바탕으로 다양한 모형 설정(변수 선택, 입지 변수의 표현, 모형의 복잡성과 깊이, 최대 변수 개수의 제한 등)에 따른 설명력의 변화를 정량적으로 비교 분석한다. 전체 표본은 임의로 두 집단(학습표본)으로 분류되어, 한 집단은 모형의 추정에 나머지 집단은 모형의 평가에 활용된다(평가표본). 이때 평가표본에 대한 설명력은 모형의 정확성과 효율성에 대한 지표로 활용이 가능하다. 이를 위해 본문은 2009년부터 2019년 사이 서울에서 거래된 아파트 620,617건을 표본으로 사용하였다. 이는 헤도

1) 즉, 이는 랜덤 포레스트 기법이 전통적 모형(헤도닉 가격 모형)에 비해 높은 예측력을 가질 수 있는 이유에 대한 분석과도 같다.

닉 모형 분석을 포함한 기존 연구들에서 사용되는 일반적인 표본의 수와 범위에 비해 상당히 큰 것으로, 표본 크기에 따른 모형의 예측력 상승을 기대할 수 있을 뿐 아니라, 본문에서 분석된 결과의 일반성을 뒷받침할 수 있을 것으로 기대한다.

이후 본문의 구성은 다음과 같다. Ⅱ장에서는 최근 증가하고 있는 기계학습 기반 모형들에 대한 국내외 선행연구들을 소개하였다. Ⅲ장에서는 랜덤 포레스트 모형이 변수를 포착하는 방식에 대하여 기술하였다. Ⅳ장에서는 Ⅲ장의 논의를 바탕으로 랜덤 포레스트와 주택 가격 결정 구조와의 관계에 대해 기술하였다. Ⅴ장에서는 주택 시장 변수의 처리 및 랜덤 포레스트 알고리즘의 설정 방식에 따라 모형의 예측력이 어떻게 변화하는지를 제시하였다. 또한, 랜덤 포레스트 기반의 모형과 전통적인 선형 헤도닉 모형과 예측력도 비교하였다. 마지막으로 Ⅵ장에서는 모형의 의의와 한계점 등을 정리하였다.

## Ⅱ. 선행연구

최근 데이터 과학 기술의 혁신을 통한 부동산 시장 연구는 크게 활성화되는 추세이다(예를 들어, Antipov and Pokryshevskaya, 2012; Čeh et al., 2018; Fan et al., 2006; Hong et al., 2020; McCluskey and Anand, 1999; Selim, 2009 등). 이는 데이터 분석 기술 자체의 혁신뿐 아니라, 데이터 수집 기술이 발전함에 따라 기계학습에 필요한 대규모의 분석 자료에 접근 가능해졌기 때문이다. 따라서 부동산 평가 모형 역시 기존의 헤도닉 가격 모형으로부터 근래에는 기계 학습이나 인공신경망에 기반한 모형으로 점차 다양해지고 있다(Wang and Li, 2019).

전통적인 헤도닉 모형은 몇 가지 가정하에서 자산의 가격을 설명변수의 회귀식으로 표현할 수 있게 해준다. 헤도닉 가격 모형의 장점은 이처럼 간단한 선형 다변량 회귀분석을 통해서 평가모형 시스템을 구축할 수 있다는 데에 있으며, 변수 간 관계를 직관적으로 이해하거나 대중에게 설명하기 쉬운 특징을 가지고 있다. 그러나, 이와 같은 단순성과 직관성은 동시에 약점이 되기도 한다. 완전 경쟁, 분리 가능한 효용구조, 완전히 통합된 시장구조 등의 엄격한 경제학적 가정이 동시에 만족되지 않는 경우, 회귀식으로 표현 가능한 단순한 함수형태는 실제의 부동산 시장이 가지고 있는 복잡성을 포착하기에 충분하지 않기 때문이다(Malpezzi, 2002; Sheppard, 1999). 결과적으로 헤도닉 모형에 기반한 부동산 평가는 그 함수형태의 단순성으로 인한 설명력의 저하 문제를 갖게 된다.

기계 학습 등에 기반한 감정평가 모형의 상당수는 변수 간 관계의 직관성을 다소 포기하는 대신, 이와 같은 설명력의 잠재적 훼손을 줄여 모형의 예측력을 극대화하려는 데에 목적이 있다. 최근 데이터 분석 기법의 폭발적 발전과 함께 다양한 데이터 분석 기법이 적용되고 있는데, 그 중에는 서포트 벡터 머신(Gu et al., 2011; Mu et al., 2014), 인공신경망 형성 기법(Limsombunchai, 2004; McCluskey and Anand, 1999; Selim, 2009), 의사결정나무(Fan et al., 2006), 랜덤 포레스트(Antipov and Pokryshevskaya, 2012;

Hong et al., 2020) 등이 있다.

서포트 벡터 머신은 분류와 회귀가 동시에 사용되는 기계 학습 모형으로, 비확률적 이진 선형 분류 모형을 형성하여 대상이 어떤 카테고리 또는 값을 가질지를 판단한다. 이를 주택 가격 예측에 적용한 사례로는 Gu et al.(2011), Mu et al.(2014), Zurada et al.(2011)이 있다. Gu et al.(2011)은 유전 알고리즘을 통해 서포트 벡터 머신의 적정 파라메터를 찾는 방식으로 중국 주택 가격의 예측 모형을 설정하였다. Mu et al.(2014)은 서포트 벡터 머신, 최소 자승 서포트 벡터 머신, 부분 최소 자승 모형을 비교한 결과, 서포트 벡터 머신과 최소 자승 서포트 벡터 머신의 예측력이 더 높다는 것을 보여주었다. Zurada et al.(2011)은 선형회귀분석, 서포트 벡터 머신, 인공신경망을 포함하는 다양한 모형들의 예측력을 비교하였다. 그 결과, 모든 경우에서 비전통적 기법이 선형회귀모형에 비해 더 나은 예측력을 제공하였으며, 특히 인공신경망과 같은 AI 기반의 모형은 분석 데이터가 비동질적일 때 다른 모형에 비해 더 나은 예측력을 제공한다는 것을 보여주었다.

인공신경망은 뇌 신경망의 작동 방식에서 영감을 얻은 통계적 학습 방식으로, 인공 뉴런의 결합 세기를 데이터 학습을 통해 변화시켜 예측력을 가질 수 있다. 이를 통한 부동산 평가 모형은 McCluskey and Anand(1999), Limsombunchai (2004), Selim(2009) 등에서 제시가 되었다. 특히 Limsombunchai(2004)와 Selim(2009)에서는 기존의 헤도닉 가격 모형과 인공신경망 기반 모형의 예측력이 비교되었는데, 인공신경방 기반 모형의 예측력이 더 우수한 것으로 나타났다.

의사결정나무는 분석 대상을 그 속성에 따른 동질군으로 분류하는 조건을 찾는 알고리즘 (Woods and Kyral, 1997)으로, 주택가격의 예측에도 사용될 수 있다. Fan et al.(2006)은 의사결정나무 방법을 통해 싱가포르 주택시장을 분석한 결과, 상당한 정도의 하부시장(sub-market)이 존재할 수 있음을 보여주었다. 이는 주택의 종류 및 특성에 따라 각 설명변수와의 관계가 다를 수 있다는 것으로, 기존의 헤도닉 가격 모형을 적용하는 경우 예측 결과의 왜곡이 생길 수 있다는 것을 의미한다.

랜덤 포레스트는 의사결정나무에 기반한 회귀모형의 앙상블 기법으로, 각 의사결정나무의 회귀결과값에 평균을 취해 얻어진다. 이는 의사결정나무 방식의 약점 중 하나인 과적합 가능성을 해소하면서도, 더 복잡한 속성 구조를 안정적으로 탐색할 수 있는 기법이다. 랜덤 포레스트는 예측력이 안정적이고 높을 뿐 아니라, 정성변수(또는 카테고리 변수)의 효과를 유연하게 처리할 수 있기 때문에 다양한 기계 학습 기법 중에서도 부동산 평가 모형 설정에 특히 적합할 수 있다(Antipov and Pokryshevskaya, 2012). 관련 연구로는 대표적으로, Antipov and Pokryshevskaya(2012), Hong et al.(2020)을 들 수 있다. Antipov and Pokryshevskaya(2012)는 랜덤 포레스트를 이용하여 아파트 가격의 평가를 시도한 초기 모형으로서, 인공 신경망과 선형회귀를 포함한 다양한 다른 기법에 비해 랜덤 포레스트가 더 효과적임을 보여주었다. Hong et al.(2020)은 랜덤 포레스트를 이용하여 서울시 강남구 아파트 가격에 대한 평가 모형을 제시하였는데, 특히 이들은 랜덤 포

레스트가 헤도닉 모형에 비해 상대적으로 뛰어날 뿐만 아니라, 오차의 평균이 5%~6% 정도에 불과할 정도로 예측력이 높을 수 있음을 보여주었다.

이처럼 기계 학습을 이용한 주택가격 예측 모형은 국내에서도 점차 활성화되어 가는 추세이다. 배성완 · 유정석(2018)은 아파트 가격지수를 예측하는 데에 서포트 벡터 머신, 랜덤 포레스트, 그래디언트 부스팅 회귀 트리, 심층신경망, LSTM과 전통적인 시계열분석 방법인 ARIMA, VECM, 베이지언 VAR, 베이지언 VECM 모형의 예측력을 비교한 결과, 기계학습의 예측력이 더 높은 것을 밝혀냈다. 이창로 · 박기호(2016)는 단독주택의 가격을 추정하는 데에 일반가산모형, 랜덤 포레스트, MARS(multivariate adaptive regression splines), 서포트 벡터머신 모형의 예측력을 사용하여 비교하였는데, 해당 연구에서는 서포트 벡터머신과 MARS의 예측력이 우수한 것으로 나타났다. 김태훈 · 홍한국(2004)은 송파구와 도봉구에서 2004년 6월에 거래된 아파트 매매 자료를 인공신경망을 통해 분석하고 회귀분석에 기반한 모형과 비교하였다. 김종수 · 이성근(2012)은 공업용 부동산의 가격 추정에 헤도닉 모형과 서포트 벡터 머신 모형을 비교 적용하였으며, 서포트 벡터 머신의 높은 예측력을 통해 기존의 헤도닉 모형의 단점을 보완할 수 있음을 제시하였다. 홍정의(2020)는 그래디언트 부스팅 기반의 세 가지 알고리즘(XGBoost, LightGBM, CatBoost)을 이용하여 서울시 아파트 가격에 대한 대량 평가를 시도한 결과, 평균 예측 오차가 극적으로 감소할 수 있음을 보여주었으며, 다양한 알고리즘을 통해 얻은 예측 모형의 합성을 통해 예측력을 증강시킬 수 있음을 제시하였다.

## Ⅲ. 랜덤 포레스트 알고리즘의 소개

### 1. 의사결정나무

랜덤 포레스트는 의사결정나무의 앙상블 기법으로, 의사결정나무의 장점은 취하고 단점은 최소화하는 기계학습 기법이다. 본문에서는 먼저 의사결정나무에 대해 소개하고, 간단히 장점과 단점을 논한 이후 랜덤 포레스트를 설명하려고 한다.

의사결정나무 알고리즘은 주어진 데이터를 분석하여 그 속에 존재하는 속성의 패턴을 예측 가능한 규칙의 구조로 나타내는 기법으로, 속성값[2]의 입력을 통해 목표로 하는 변수의 값 또는 상태를 예측하는 모형을 생성하는 것을 목표로 한다.

의사결정나무는 규칙노드들과 리프노드들로 이루어져 있다. 규칙 노드는 어떤 속성에 대한 조건문을 의미하며, 데이터가 그 조건에 부합하는지 아닌지를 판단하여 분류한다. 리프 노드는 이러한 조건들의 구조에서 최종적으로 분류된 결정값을 의미한다.

의사결정나무는 최종적으로 분류된 데이터들이 가장 균일한 정보량을 갖는 규칙노드의 구조를

2) 일반적인 경제학 문헌에서는 각 데이터 표본이 갖는 성질들을 변수라고 표현하는데, 의사결정나무에서는 예측 대상이 되는 변수와 그것을 예측하기 위해 관찰하는 변수를 구분하여, 예측 대상 변수의 상태 또는 값을 결정값 그 외 변수의 값을 속성값으로 지칭하는 경향이 있다.

 https://doi.org/10.30902/jrea.2021.7.1.1

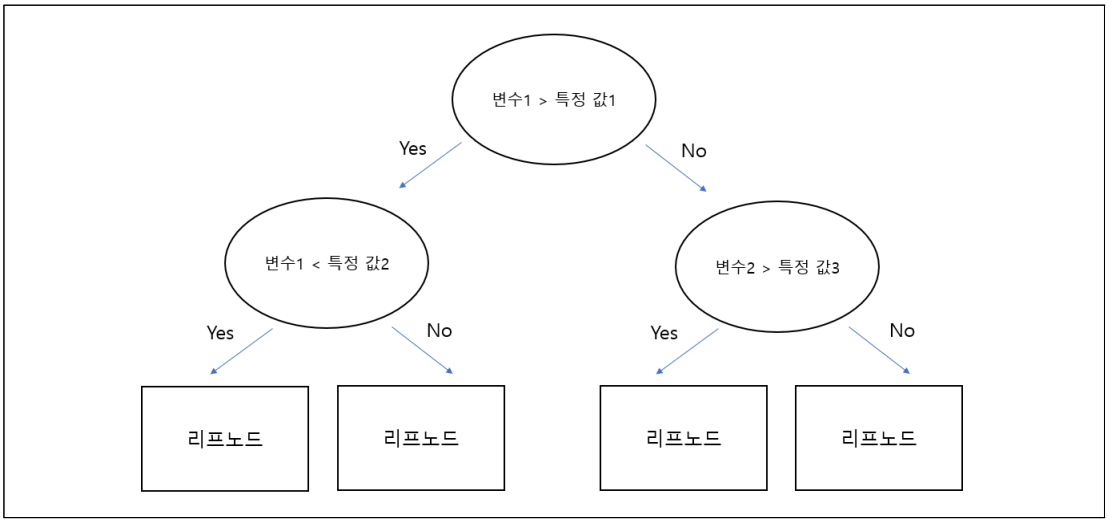
찾는 알고리즘이다. 즉, 의사결정나무 알고리즘은 데이터를 분할하는 데 가장 좋은 조건을 반복적으로 찾아 규칙 노드의 구조를 생성한다. 이때 정보의 균일도를 평가하는 방식은 의사결정나무를 통해 풀고 있는 문제가 분류인지, 수치예측(회귀)인지 등에 따라 차이를 보일 수 있다. 분류 문제의 경우, 정보획득량(information gain), 지니계수(gini index), 엔트로피 지수(entropy index)를 사용하며, 회귀문제의 경우 일반적으로 리프 노드에서의 분산값을 사용한다.

모형의 설계자는 하부 규칙 노드의 최대 개수 및 규칙 노드 분할에 필요한 최소 데이터 개수, 리프 노드로 편성될 수 있는 최소 데이터 개수 등을 하이퍼 파라미터로 설정할 수 있다. 설정된 하이퍼 파라미터 하에서, 더 이상 규칙 노드를 형성할 수 없으면 규칙 분할을 멈추고 구조를 결정한다. 〈그림 1〉은 의사결정나무의 구조를 개략적으로 보여주고 있다.

그러므로 이는 회귀뿐 아니라 분류에도 사용될 수 있다. 다만 분류의 경우, 예측 대상이 카테고리 변수일 때 사용되기 때문에, 본문과 같이 연속 변수인 아파트 매매가격을 평가하는 모형을 만드는 경우 회귀 분석을 수행해야 한다. 분류 문제의 경우, 입력된 속성을 바탕으로 최종 분류된 노드의 최빈값을, 회귀 문제의 경우 리프에 속한 데이터의 평균값을 반환한다. 이때 예측값 종류의 개수는 리프 노드의 개수보다 클 수 없다.

## 2. 랜덤 포레스트와 의사결정나무

의사결정나무는 변수 단위의 설명력을 유지하는 직관적이고 강력한 알고리즘이다. 또한, 정보의 균일도에 기반하여 규칙을 생성하므로, 속성값의 전처리가 복잡하지 않은 장점도 가지고 있다. 하지만, 동시에 과적합에 대한 위험에 노출되어 있다. 설정된 최대 깊이, 최소 분할 기준, 최소

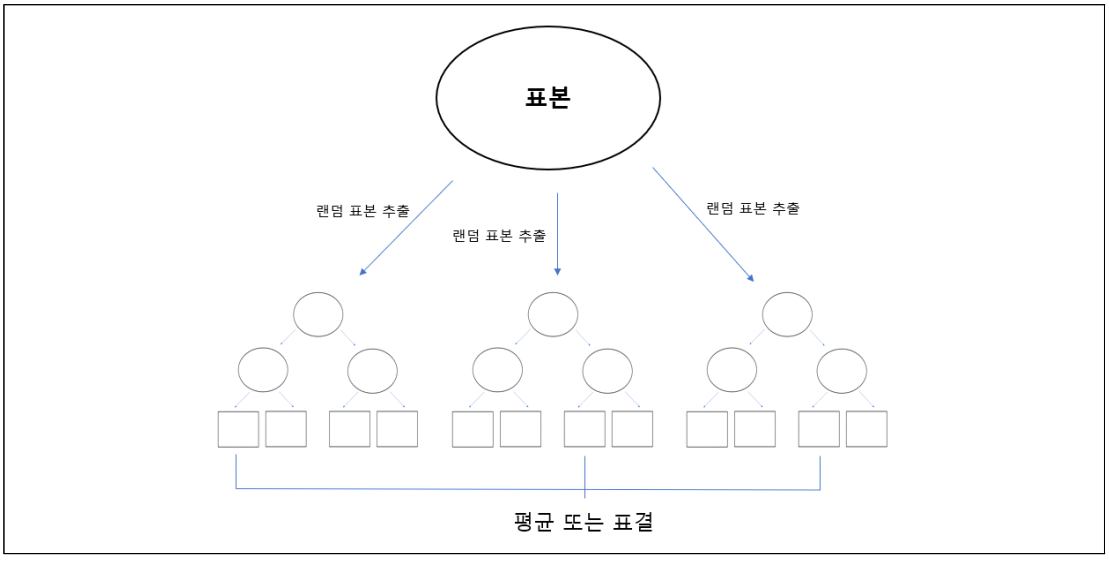



〈그림 1〉 의사결정나무의 개념

리프 노드 등에 따라, 데이터의 속성을 분할하는 규칙 구조가 더 이상 없음에도 규칙 노드가 계속 추가되는 경우를 생각해 볼 수 있다. 노드의 하단으로 갈수록 표본 집단의 크기가 작아지게 되므로, 이런 경우 학습 집단의 속성이 가진 소음(noise)이 새로운 규칙 노드를 만드는 데에 크게 영향을 미칠 수 있으며, 결과적으로 학습 데이터에 대한 과적합으로 이어질 수 있다. 이러한 과적합이 생기는 경우, 학습 데이터에 대한 예측력은 향상되나, 평가 데이터 또는 일반 데이터에 대한 예측 성능은 오히려 저하된다.

이러한 단점은 극복하고 의사결정나무가 가진 직관성과 예측력을 최대한 살린 기법이 랜덤 포레스트 알고리즘이다. 랜덤 포레스트는 부트스트래핑(bootstrapping)을 통해 생성된 다수의 의사결정나무가 보팅을 통해 최종 예측을 하는 기법이다. 전체 표본에서 추출된 많은 임의 표본들을 대상으로 수많은 의사결정나무를 계산하고, 각각의 의사결정나무에서 반환된 값을 평균 또는 표결(보팅)로 합성하는 것이다. 〈그림 2〉는 이러한 랜덤 포레스트의 계산 방식을 도식으로 나타내고 있다.

분류 문제에서 보팅의 방법에는 두 가지가 있으며, 하드 보팅과 소프트 보팅으로 나뉜다. 하드 보팅은 다수결의 원칙과 유사한데, 예측된 값 중 다수가 결정한 값을 최종 결과로 제시하는 것이다. 소프트 보팅은 각 결정값의 결정 확률을 모두 더한 뒤 평균을 취해 가장 확률이 높은 결정값을 제시하는 것이다. 회귀 문제의 경우, 각 의사결정나무에서 반환된 결정값의 평균을 제시한다. 주택가격 예측의 문제는 회귀 문제에 해당하므로, 가장 단순한 소프트 보팅에서는 언급한 바와 같이 각 의사결정나무 반환 값의 평균을 사용한다.




〈그림 2〉 랜덤 포레스트의 개념

 https://doi.org/10.30902/jrea.2021.7.1.1

## Ⅳ. 랜덤 포레스트와 주택 시장

앞에서는 랜덤 포레스트의 일반적인 작동 원리에 대해 개략적으로 서술하였다. 이 장에서는 이러한 알고리즘이 전통적인 헤도닉 모형에 비해 주택 가격 예측에 있어 더 높은 정확성을 가지는 이유를 고찰하고, 주택 대량평가에의 활용에 있어 랜덤 포레스트 효율성을 극대화 시키기 위해서는 어떤 설계가 필요한지를 논의한다.

### 1. 하위시장과 랜덤 포레스트

주택 시장의 본질적인 특성 중의 하나는 하위시장의 존재이다(김경환·손재영, 2010). 주택 시장은 하나의 완전한 동질적인 시장으로 존재하지 않으며, 유형/규모/품질/입지에 따라 구별되는 작은 다수의 시장으로 구성된다(김주영 · 우경, 2004; 정건섭 · 이상엽, 2007; Watkins, 2001 등). 예를 들어, 대형 주택시장에서는 가족원의 수가 상대적으로 많을 것이므로, 화장실의 개수가 가격에 뚜렷하게 영향을 미칠 수 있다면 소형 주택시장에서는 화장실의 수가 1개를 초과하는 경우, 가격에 영향을 미치지 못할 수 있다. 하위시장은 주택 특징이나 크기별로만 존재하는 것이 아니라, 입지에 따라서도 분화할 수 있다(손철, 2011).

하위시장의 존재는 헤도닉 가격 모형의 설계단계에서 고려해야 하는 중요한 요소이며, 때로 예측력의 훼손이 일어나는 원인이기도 하다(김경환·손재영, 2010). 왜냐하면 하나의 하위시장에서 성립되는 가격의 구조는 다른 하위시장과 다를 수 있기 때문이다. 헤도닉 가격 모형의 특징은 선형 다변량 회귀분석에 적합한 형식으로 자산 가격과 그 속성의 관계를 표현하고 있다. 이러한 특성상, 어떤 변수가 가격에 미치는 영향은 하나의 계수 값으로 표현된다. 그러나 하부 시장이 존재하는 경우 어떤 속성이 가격 결정 구조에 미치는 영향이 분절된 하위시장에 따라 다를 수 있기 때문에, 결과적으로 어떤 하위시장을 잘 설명하는 헤도닉 가격함수가 다른 하위시장에는 잘 적용되지 않을 수 있다.

한편, 랜덤 포레스트 예측의 바탕이 되는 의사결정나무 기법은 주택 시장에 존재하는 자산들(분석 대상인 데이터 집합 전체)을 유사한 특성을 갖도록 분류하는 위계적 조건의 구조를 찾아낸다(Fan and Koh, 2006). 이때, 특정 조건 하부의 조건 구조는 다른 조건의 하부 조건 구조와 동일할 필요가 없다는 것이 중요하다. 즉, 상부의 규칙 노드에 의해 다른 노드로 구분되는 경우, 그 이후로는 다른 조건의 구조에 따라 분류될 수 있다는 것이다.

랜덤 포레스트는 의사결정나무의 앙상블 기법이므로 그와 동일한 특성을 가지고 있다고 볼 수 있다. 랜덤 포레스트 모형에서 각 규칙 노드는 주택 시장에서의 어떤 변수(예를 들어, 주택의 크기, 방 개수, 입지 등)에 대한 조건을 의미한다. 이는 랜덤 포레스트를 통해 생성된 조건의 위계적 연쇄 구조는 하부 주택 시장의 구조 자체를 각각 묘사하는 방법이 될 수 있음을 의미한다. 예를 들어, 상부 규칙 노드에서 주택의 크기가 일정 수준 이상인 경우와 이하인 경우로 나뉘었다면, 크기가 큰 주택의 가격 결정 구조와 작은 주택은 그 가격 결정 구조(그 구분된 하위시장의 특징에 맞추어)

가 다르게 묘사될 수 있다는 것이다. 이와 같은 특징에 따라, 랜덤 포레스트 알고리즘은 헤도닉 모형과 달리 하위시장이 존재하더라도 각 하위시장의 가격 결정 구조에 따로 포착할 수 있으며, 결과적으로 설명력의 훼손을 방지할 수 있다.

중요한 것은, 주택 하위시장의 구조가 상당히 복잡하다는 것이다(김경환·손재영, 2010). 왜냐하면 하위시장은 기본적으로 서로 연계를 가지면서도 분리된 성격을 가지기 때문이다. 예를 들어, 서울 주택시장을 조망하는 경우, 하위시장은 고가주택/저가주택 간이나 대형주택/소형주택 간에도 성립할 수 있고, 입지별로도 성립할 수 있다. 이때 각 하위시장은 분리되었으면서도 명백히 그 각 분류 안에서는 연계성을 가지고 있으므로, 예측 모형의 정확도를 향상시키기 위해서는 하위시장 간 위계구조를 모형이 효과적으로 포착할 수 있어야 한다.

이때, 랜덤 포레스트는 정보량이 높은 변수 순으로 규칙 노드를 형성하므로, 주택 시장과 같이 다양한 하위시장이 연결된 경우, 그 하위시장 간의 위계를 데이터 기반으로 찾는 알고리즘의 역할을 한다. 이런 경우, 예측 모형은 분석 대상이 가진 구조의 복잡성과 표본수의 확보[3]를 동시에 포착할 수 있다. 즉, 랜덤 포레스트 모형의 활용은 주택시장과 같이 하위시장이 복잡한 경우 모형의 설명력을 높이는 데에 크게 기여할 수 있다.

이런 경우, 랜덤 포레스트 기반 모형이 높은 예측력을 갖기 위해서는 실제 시장의 복잡성을 반영하기에 충분한 만큼의 논리적 깊이와 특이성을 적절히 포착할 수 있어야 한다. 이는 랜덤 포레스트 모형에서 규칙 노드의 최대 깊이와 리프가 되기 위한 최소 데이터의 수를 결정하는 파라메터의 설정과 연관되어 있다. 특히 리프를 구성하는 최소 데이터의 수는 이 맥락에서 가장 작은 하위시장(동질성을 갖는 가장 작은 단위)으로 분류될 수 있는 최소 단위를 의미하므로, 이는 실제의 특이성과 복잡성을 포착할 수 있는 만큼 복잡하게 주어져야 설명력 훼손이 일어나지 않을 것이다.

한편, 이는 고려되는 변수의 숫자와도 연관되어 있다. 만약 실제 주택시장의 가격 결정 구조에 영향을 미치는 속성들에 비해 너무 많은 속성들이 연산 과정에서 고려되고 있다면, 오히려 이로 인한 예측 성능의 저하가 생길 수 있다. 예를 들어, 실제로는 화장실의 개수가 주택 가격에 영향을 미치지 못한다고 해도, 결정나무의 하부 노드에서 샘플 데이터 수가 충분히 작아지는 경우, 같은 노드에 묶인 데이터에 포함된 소음의 편향성에 의해 화장실의 개수에 대한 규칙 노드가 생성될 수 있다. 이는 일종의 표본편향으로 인한 과적합이므로, 결과적으로 모형의 예측 성능 저하로 이어질 것이다. 그러므로, 랜덤 포레스트를 통해 주택 가격 모형을 설정할 때에는 속성이 불필요하게 많이 포함되지 않도록 변수 개수의 조절을 하는 것이 예측 성능을 높이는 데에 도움이 될 수 있다[4].

또한, 이는 정성 변수의 표현 방식과도 간접적으로 연관될 수 있다. 정성 변수의 경우, 일반적으

3) 만약 헤도닉 모형을 추정하기 위해 가능한 모든 하위시장을 완전히 분리된 시장으로 가정한다면 적은 표본수로 인한 예측력 저하를 겪게 되며, 반대로 하위시장을 하나의 시장으로 가정한다면 구조의 단순성으로 인한 예측력 저하를 겪게 된다.

4) 이와 관련한 속성의 적정 개수, 변수 처리방식, 최대 깊이 및 최소 리프 분할 설정 등의 정량분석은 Ⅴ-2절에서 보다 자세히 다루었다.

 https://doi.org/10.30902/jrea.2021.7.1.1

로 헤도믹 모형에서는 더미변수로 표현되나, 랜덤 포레스트 모형에서 더미변수화 되거나 라벨변수화(즉, 각 범주별로 하나의 고유한 정수를 부여받는 것) 될 수 있다. 이때 정성변수의 더미화 과정에서는 그 변수의 범주 수만큼의 변수가 모형에 포함될 것임을 알 수 있다. 그런데 변수 개수가 늘어나는 만큼 연산 효율성뿐 아니라, 과적합 가능성도 늘어나게 되므로, 결과적으로 모형의 설명력을 낮출 가능성이 있다.

## 2. 입지가치: 비선형성의 처리방식

부동산 가격에 영향을 미치는 요소는 크게 두 가지로 분류할 수 있다. 첫째는, 건축물 자체의 속성인 구조적 속성으로, 건물의 크기, 방 수, 층 수, 난방 특징 등이 이에 속한다. 이러한 속성들은 상대적으로 관찰 및 측정하기가 용이하며, 가격에 미치는 영향 역시 직관적이다. 둘째는, 해당 자산의 위치의 가치인 입지적 속성이다. 이는 주요 시설로부터의 거리나 이웃 특성 등 주택의 입지 가치에 영향을 미치는 모든 요소를 의미한다(김경환 · 손재영, 2010).

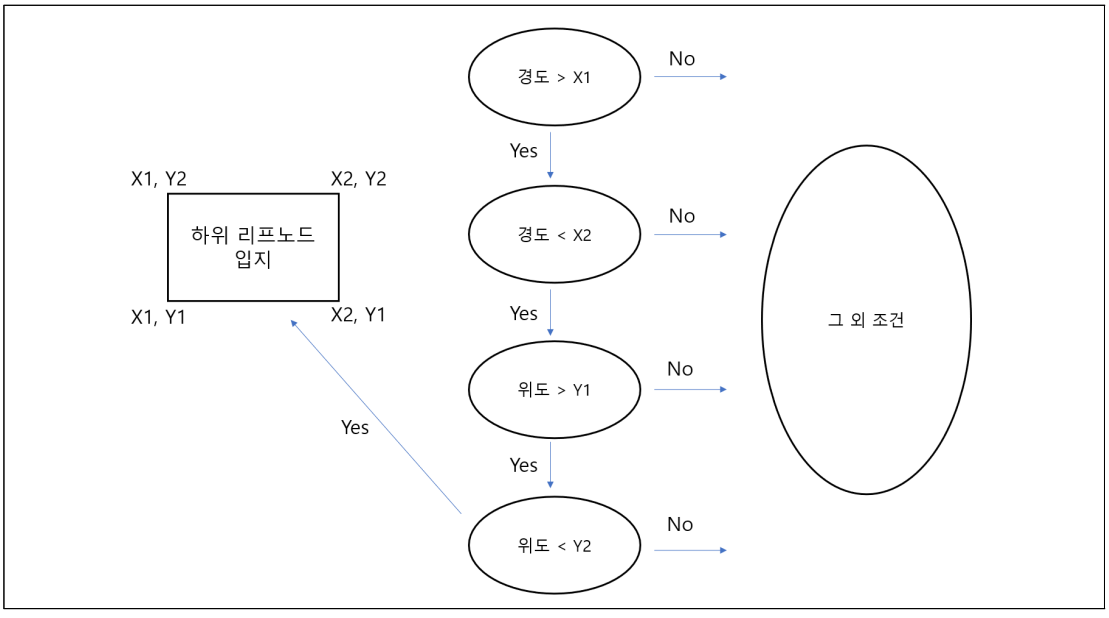
한 가지 문제는 이러한 입지적 속성들은 구조적 속성에 비해 관찰 가능한 변수들로 포착하기에도 변수 간의 관계를 모형으로 표현하기도 훨씬 어렵다는 것이다(Adair et al., 1996; Heyman and Berghauser, 2019). 헤도닉 모형에서는 주택의 시장 가치와 그에 영향을 주는 속성을 일반적으로 단순한 선형 관계식으로 표현하는데, 이는 모형 예측력 저하의 원인 중 하나가 될 수 있다. 예를 들어 다른 시설에 대한 접근성을 생각해 보자. 다른 시설에 접근성이 떨어질수록(즉 이동비용이 증가할수록) 비효용이 증가하므로, 시장이 경쟁적이라면 각 입지는 그러한 비효용을 낮춰주는 것만큼의 시장 가치를 지니게 된다(Alonso, 1964). 그런데 주요 시설까지의 이동비용 자체를 직접 관찰하는 것은 불가능하므로, 일반적으로 헤도닉 모형에서는 먼저 각 자산의 입지별로 주요 시설들을 정의한 다음, 그 시설로부터의 물리적 거리(가장 단순한 경우, 2차원 평면 상에서의 유클리드 거리)가 이동비용에 비례한다는 가정하에 시장 가격과 각 시설로부터의 관계식을 세우게 된다. 그러나, 각 주택의 입지별 주요 시설은 완전하게 정의될 수도 없거니와 각 시설과의 유클리드 거리가 실제 이동비용 함수를 완벽히 묘사한다고 볼 수도 없다(Adair et al., 1996). 이런 경우, 선형 회귀 모형의 가격 예측 성능은 실제 이동비용을 단순화한 만큼 저하될 것이다. 접근성 변수는 입지 속성 중 상대적으로 용이하게 관찰할 수 있는 변수에 해당하며, 실제 입지 가치는 보다 복잡하게 형성된다. 예를 들어, 주택 가치는 이웃으로부터의 외부효과(Nesheim, 2002)나 문화적 특성(Sheppard, 2010; Zukin, 1987), 지역 노동시장의 특성(Osland and Thorsen, 2008; Wheaton and Lewis, 2002)에 의해 영향을 받을 수 있다. 이러한 변수들은 그 영향력을 제대로 측정하기 어려운 형태로 존재하므로, 변수 간 관계를 함수의 형태로 직접 묘사하는 헤도닉 모형에서는 예측력의 훼손을 일으키는 주요한 요인 중 하나로 작용할 수 있다.

반면, 랜덤 포레스트 알고리즘은 가격 예측에서 주택 가치와 속성 간의 명시적 관계식을 필요

로 하지 않는다. 랜덤 포레스트는 규칙 노드의 구조를 통해 전체 표본을 분류하여 그 최종 리프의 평균값을 반환한다. 즉, 규칙 노드를 형성하는 단계에서 속성값들이 이질적인 표본 간의 차이를 적절히 구분할 수 있는 형식으로만 짜여져 있으면, 그 속성과 시장가치 간에 어떤 관계가 있는지를 고려하지 않더라도 무방하다.

이러한 특징은 주택 가격에 대한 예측 과정에서 입지 효과를 포착하는 데에 효율적으로 적용될 수 있다. 〈그림 3〉은 랜덤 포레스트(결정 나무) 알고리즘이 규칙 노드를 생성하는 단계에서 위도와 경도를 어떻게 이용할 수 있는지에 대한 예시를 보여준다. 위도와 경도의 경우, 헤도닉 모형에서는 직접적으로 사용될 수 없으나, 랜덤 포레스트에서는 예시와 같이 적용될 수 있다. 그림을 보면 알 수 있듯이, 위도와 경도에 대한 규칙 노드의 연쇄는 해당 입지에 대한 그리드(grid)를 형성할 수 있다. 그러한 조건문 하위로 분류된 모든 주택은 동일한 입지 그리드에 위치한 주택이므로, 결과적으로 해당 입지 그리드의 평균적 입지효과를 포착한 결과를 반환할 수 있다.

이는 랜덤 포레스트가 입지적 측면에서 동질적인 것으로 판단되는 가장 작은 단위의 그룹(즉, 가장 세밀하게 분류된 하위시장)의 평균적 성질을 통해 예측을 수행한다는 것을 의미한다. 즉, 주택 가치와 입지효과 간의 명시적 함수 관계를 정의하지 않고 예측을 수행한다. 이 경우, 비록 어떤 입지 효과들이 얼마나 그 가치에 영향을 미치는지에 대해서는 알 수 없지만, 예측된 최종적 주택 가격에는 해당 입지 가치에 영향을 미치는 모든 효과(관측되지 않는 효과를 포함하여)가 이미 반영된 것으로 해석할 수 있다. 입지효과를 일으키는 변




〈그림 3〉 랜덤 포레스트 하에서의 입지 분류

 https://doi.org/10.30902/jrea.2021.7.1.1

수와 함수형식을 특정하지 않음으로써, 오히려 직접 관측할 수 없는 차이를 포착할 수 있는 것이다(Hong et al., 2020).

이러한 종류의 예측 성능은 입지 효과를 적절히 포착하기에 충분한 수의 표본(특히, 공간 집약적인 표본)을 확보하는 데에 따라 크게 달라질 수 있다(Hong et al., 2020). 만약 충분한 표본을 확보했다면, 헤도닉 모형이 포착하지 못한 입지 효과의 복잡성의 크기만큼 랜덤 포레스트의 예측 성능이 더 우수하게 나타날 수 있다(Antipov and Pokryshevskaya, 2012).

이처럼 랜덤 포레스트가 입지가치를 포착하는 방식을 고려하면, 랜덤 포레스트가 예측력뿐 아니라 모형 추정에 요구되는 정보량의 효율성 측면 역시 높다는 것을 알 수 있다. 즉, 랜덤 포레스트에서는 입지 효과와 주택 가치 간의 관계가 특정한 함수로 표현되지 않으므로, 헤도닉 모형에서 입지 가치를 포착하기 위해 사용하는 전통적인 변수들은 랜덤 포레스트 모형에서는 사용할 필요가 없다는 것이다. 각 주택의 위치를 포착할 수 있는 정보인 좌표값(위도, 경도)만 있더라도 랜덤 포레스트 모형은 복잡한 입지 가치의 차이를 직접 포착할 수 있다. 만약 표본 수가 충분하고 모형이 제대로 설계되었다면, 좌표값 이외의 입지변수들을 추가로 모형에 포함했다고 해도 모형 설명력의 상승이 제한될 것임을 예측할 수 있다.

## 3. 거시경제환경과 주택시장

주택 시장은 금융 시장, 지역 노동 시장, 경제성장률 등 거시 경제적 환경과 긴밀하게 연결되어 있다. 헤도닉 모형에서는 다기간에 걸친 자료를 사용할 때 이러한 거시경제환경에서의 차이를 주로 더미 변수를 사용하여 통제한다. 그러나 이는 거시경제 환경의 영향을 특정년도에 거래된 주택에 나타나는 평균적 가격 효과로 묘사하는 것으로, 모든 주택에 완전히 동일한 거시경제효과가 나타난다는 것을 의미한다. 그러므로, 만약 거시경제 환경의 변화(헤도닉 모형 내에서 단일 시장군으로 포함하고 있는 범위 내에서)가 표본의 특성에 따라 다른 효과를 갖는다면, 그러한 복잡성만큼의 설명력 훼손이 나타나게 된다.

또한, 선형회귀모형의 특성상 일반적으로 거시경제환경의 효과는 다른 속성과 주택 가격의 관계와 독립적으로 묘사된다. 그러나 실제는 거시경제 환경은 주택의 종류에 따라 다른 효과가 나타날 수 있으며, 때로는 변수와 주택 가격의 관계 자체에 영향을 미칠 수도 있다. 대표적인 이유로, 거시적 소득 변화가 생기는 경우 대체 효과가 생기기 때문이다. 주택 자체는 필수재이나, 주택가격모형(특히, 헤도닉 모형)에서 중요한 것은 주택의 각 속성에 대한 수요이다. 이때 주택의 어떤 속성, 예를 들어 주택의 점유공간에 대한 소득 탄력성은 주택의 입지에 대한 소득 탄력성보다 크다고 하면, 결과적으로 일반적 소득 변화는 다른 속성을 갖는 주택 간에 다른 효과를 낼 수밖에 없을 것이다.

이러한 복잡성들은 랜덤 포레스트 하에서 효과적으로 포착될 수 있다. 앞 절에서도 언급된 바와 같이 랜덤 포레스트에서는 하위 노드에서는 다른 하위 노드와 전혀 다른 변수 관계를 정의할 수 있기 때문이다. 즉, 랜덤 포레스트에서는 거시경제적 변화가 나타나는 각 시점별로 속성과 가격의

관계를 구조화할 수도 있으며, 표본의 특성에 따라 다른 효과도 포착할 수 있다는 것이다. 특히, 랜덤 포레스트는 거시경제효과 자체를 묘사하는 모형이 아니라, 시점에 따라 달라진 변화를 데이터 기반으로 감지하여 그 차이를 구분하는 모형이므로, 명시적으로 어떤 거시 경제환경 변화가 있는지를 변수화할 필요 없이 단순히 시점 구분만 하는 것으로 충분하다. 다른 말로 하면, 랜덤 포레스트 모형에서는 시점을 구분하는 변수를 포함하는 것으로 충분하며, 그 외의 거시경제변수(금리, 경제성장률 등) 등을 직접적으로 고려할 필요가 없다는 것이다. 이는 입지효과 포착의 경우와 유사하다. 또한, 이는 연구자가 사전에 자료의 동질성을 해치지 않는 범위의 시점을 설정하지 않아도 모형의 설명력이 높게 유지될 수 있음을 의미한다.

## Ⅴ. 정량분석

앞에서는 부동산 시장과 랜덤 포레스트를 통한 가격 예측 간의 관계에 대하여 논의하였다. 이 절에서는 랜덤 포레스트를 통한 주택 평가 모형을 정량적으로 분석하여, 보다 설명력이 높은 모형을 최소한의 정보로 구성하기 위한 모형의 설계에 대해 논의한다. 특히, 파라메터 및 변수의 표현 등이 모형의 정확성에 미치는 효과를 검사하고, 헤도닉 모형과의 예측력을 비교하려고 한다.

### 1. 자료 및 변수 설명

정량분석을 위해 본 연구는 국토교통부, 통계청, 네이버 부동산 그리고 구글맵을 이용하여 주택 시장 데이터를 수집하였다. 분석 대상은 2009년부터 2019년도까지 서울시에서 일어난 아파트 거래이다. 아파트는 단독주택에 비해 상대적으로 데이터로 집계되는 주택 속성(층, 크기, 시설 등)이 균일하기 때문에, 관찰 가능한 주택 속성을 통해 가격을 예측해야 하는 본문의 분석의 특성상 더 적합한 분석 대상일 수 있다. 총 표본 수는 620,617건으로, 지금까지 이루어진 부동산 가격 분석 사례(특히, 헤도닉 가격 모형) 중에서도 상대적으로 방대한 자료를 대상으로 하고 있다.

본문의 정량분석에 사용된 변수의 설정은 지금까지 다양하게 이루어진 헤도닉 모형 연구에 종합적으로 기반하되, 특히 우리나라 수도권 아파트 시장을 대상으로 한 연구들(구본창 · 손영현, 2001; 김민성 · 박세운, 2014; 김우성 외, 2019; 김운정, 2004; 김천일, 2018; 손아남 · 정경수, 2014; 원두환 · 김형건, 2008; 윤채규, 2003; 이강 · 최근희, 2016; 조주현, 1998; 정수연 · 김태훈, 2007; 하유정 · 이현석, 2020; Hong et al., 2020 등)을 폭넓게 참고하였다.

가장 기본적인 변수인 매매가격은 국토교통부에서 공개하는 아파트 매매 실거래가를 사용하였다. 주택 속성은 구본창 · 송영현(2001), 김우성 외(2019), 윤채규(2003), Hong et al.(2020)과 같이 구조적 특성, 환경적 특성, 입지적 특성으로 구분하여 나누어 수집하였다. 구조적 속성은 아파트 건축물 자체의 다양한 특징을 의미하며, 일반적으로 헤도닉 모형에서는 전용 면적, 방 수, 화장실 수, 경과 연수를 공통적으로 포함한다(김우성 외, 2019; 윤채규, 2003; 조주현, 1998 등).

 https://doi.org/10.30902/jrea.2021.7.1.1

본문에서는 이에 더해 난방 시스템, 복도 형식, 층수, 경과 연수를 분석에 사용하였다. 난방 시스템의 경우, 원두환 · 김형건(2008)에서 알려진 바와 같이 주택 가격에 영향을 미치는 요소로 고려할 수 있다. 본문에서는 이를 3가지 난방방식(개별난방, 중앙난방, 지역난방)을 분류하는 카테고리 변수로 설정하였다. 복도 형식은 김우성 외(2019), Hong et al.(2020)에서 고려되었으며, 난방 방식과 유사하게 3가지 복도 형식(복도식, 계단식, 혼합식)을 구분하는 카테고리 변수로 표현되었다. 아파트 층수에 대한 고려는 정수연 · 김태훈(2007)의 연구에 의해 뒷받침되었다. 이는 해당 아파트의 층 수를 정수로 표현하는 정량 변수이며, 지하가 있는 경우 음수로 표현되었다.

환경적 특성은 아파트 단지 전체의 특성을 의미한다. 이는 맥락에 따라 구조적 특성으로 분류되기도 하지만, 본문에서는 구본창 · 송영현(2001), 김우성 외(2019), 윤채규(2003) 등에 따라 구조적 특성과 별개의 단지 특성으로 분류하였다. 본문에서는 환경적 특성으로 아파트 단지의 크기(총 세대 수, 동 수), 평균 주차 가능대수, 단지 내 최고층의 높이, 최저층의 높이, 건폐율, 용적률이 고려되었다. 아파트 단지의 크기는 단지의 총 세대 수 또는 동 수로 표현되며, 도시 경제학에서 의미하는 군집 효과(Glaeser, 2019; Tabuchi, 1998)의 크기를 의미하는 것으로 볼 수 있다. 이처럼 단지의 크기를 고려한 헤도닉 모형으로는 김덕중(2002), 김운정(2004), 김진유 · 이창무(2005), 최윤아 · 송병하(2006) 등을 찾을 수 있다. 평균 주차공간의 경우, 이강 · 최근희(2016)의 연구에서도 알 수 있듯이, 최근 들어 가구당 주차수요가 증가하면서 점차 중요한 요인으로 고려된다. 본문에서 이는 단지 내 총 주차 가능 대수를 세대수로 나눈 것을 의미한다. 최고층과 최저층은 단지 내의 조경 등과 연관이 있다. 윤정중(2001)에서도 알 수 있듯이, 아파트 단지의 가치는 조명경관에 의해서도 영향을 받을 수 있는데, 이때 단지 내 각 건물의 최고층수는 단지의 경관에 영향을 미칠 가능성이 높다(정수연 · 김태훈, 2007 등을 참고). 건폐율과 용적률은 각각 건축면적에 대한 대지면적의 비율과 건축물 총면적의 대지면적에 대한 비율을 의미한다. 김우성 외(2019), 김천일(2018) 등에 따르면, 건폐율과 용적률은 재건축 시 기대할 수 있는 수익과 직접적인 연관이 있기 때문에 아파트 가격에 영향을 미치는 요소로 고려되었다.

입지적 특성은 일반적으로 헤도닉 모형에서는 주요 시설(특히 지하철이나 학교 등)과의 거리로 표현되는데, 이는 헤도닉 모형의 함수 표현 특성상 가격에 영향을 미치는 요소를 측정 가능한 형식으로 직접 변수화해야 하기 때문이다. 그러나 랜덤 포레스트에서는 이와 같이 가격에 비례하는 영향을 미치는 변수를 명기화 해야 할 이유가 없을 뿐 아니라, 그런 경우 오히려 직접 관측하기 힘든 요소의 영향력을 단순화할 수 있다. 따라서 본문에서는 전통적인 헤도닉 모형에서 고려하는 형식의 접근성 변수들뿐 아니라, 좌표 자체를 수집하여 그 정량분석 결과를 비교 · 분석하였다. 먼저, 전통적인 접근성 변수로는 해당 아파트로부터 가장 가까운 지하철, 초등학교, 중학교, 고등학교, 대학교, 근린 공원, 박물관, 행정센터와의 거리(m)를 고려하였다. 이는 주택 가격에 교육 환경

(하유정 · 이현석, 2020), 인근 공원(손아남 · 정경수, 2014), 지하철 접근성(김민성 · 박세운, 2014)이 영향을 미칠 수 있다는 기존 연구들에 기반하여 설정되었다. 이러한 접근성 변수들은 모두 구글 맵 상의 위도, 경도 값을 통해 각 아파트와의 유클리드 거리로 측정되었다. 또한, 입지변수의 하나로 행정동을 고려하였다. 이는 각 행정동의 이름을 값으로 갖는 카테고리 변수이다. 이는 넓은 지리적 범위를 대상으로 한 공간분석에서 상당한 설명력을 갖는 변수로 알려져 있다(김우성 외, 2019). 마지막으로, 각 아파트의 위도와 경도 값 자체를 수집하였다. 헤도닉 모형의 경우, 위도와 경도 자체를 직접 모형에 반영할 수 없으나, 랜덤 포레스트의 경우 이를 직접 예측 과정에서 사용할 수 있다. 이러한 변수들의 기초통계는 〈표 1〉에 나타나 있다.

〈표 1〉 기초통계

| 변수 | 단위 | 평균 | 표준편차 | 최소 | 최고 |
|---|---|---|---|---|---|
| 가격 | 원 | 54,043.9 | 36,949.5 | 700 | 700,000 |
| 크기 | 평방미터 | 79.5 | 28.5 | 12 | 325.4 |
| 방수 | 개 | 3 | 0.7 | 1 | 8 |
| 화장실수 | 개 | 1.7 | 0.5 | 1 | 5 |
| 아파트 세대 수 | 세대수 | 1,032.1 | 1,153.4 | 5 | 9,510 |
| 아파트 단지 수 | 단지수 | 11.9 | 14 | 1 | 122 |
| 평균 주차가능대수 | 개수 | 1.1 | 0.5 | 0 | 12 |
| 용적률 | % | 287.9 | 126.9 | 2 | 1,477 |
| 건폐율 | % | 24.9 | 27.6 | 2 | 2,457 |
| 최고층수 | 층수 | 19.2 | 6.8 | 4 | 69 |
| 최저층수 | 층수 | 12.2 | 5.4 | 1 | 54 |
| 층수 | 층수 | 9.4 | 6.2 | -4 | 68 |
| 경과 연수 | 연수 | 13.8 | 7.8 | 0 | 49 |
| 지하철로부터의 거리 | 미터 | 770.6 | 618.1 | 2.6 | 5,583.6 |
| 공원으로부터의 거리 | 미터 | 1,036.2 | 526.2 | 55.7 | 3,268.2 |
| 초등학교로부터의 거리 | 미터 | 337.2 | 169.5 | 10.6 | 1,810 |
| 중학교로부터의 거리 | 미터 | 471.2 | 252.6 | 2.6 | 2,130.2 |
| 고등학교로부터의 거리 | 미터 | 579.9 | 333.7 | 24.6 | 2,837.4 |
| 대학교로부터의 거리 | 미터 | 1,868 | 1,190.6 | 50.3 | 7,111.5 |
| 박물관로부터의 거리 | 미터 | 1,829.8 | 1,068.6 | 35.3 | 6,839.1 |
| 행정센터로부터의 거리 | 미터 | 1,938.6 | 992.8 | 16.8 | 6,521.7 |

 https://doi.org/10.30902/jrea.2021.7.1.1

## 2. 결과 분석

### 1) 모형의 설명력 측정방법

대량평가에 있어 중요한 것은 모형의 설명력이다. 실제 시장의 가격결정구조를 모형이 보다 자세한 부분까지 포착하는 경우(즉, 설명력이 높은 경우) 결과적으로 그 예측의 정확성이 상승하게 된다. 이때 성과측정에 있어 염두에 두어야 하는 것은 표본 집합의 분할이 필요하다는 것이다. 만약 모든 표본을 사용하여 모형을 측정하는 경우, 모형이 그 자료에 과적합을 일으키더라도 이를 판별할 수가 없다. 과적합은 모형의 예측 방향이 학습 표본(모형을 추정하는 데에 사용된 표본)에 필요 이상으로 맞추어져, 결과적으로 학습 표본에 존재하는 특이성을 모형의 예측에 반영하게 되는 것이다. 이러한 과적합 문제가 나타나는 경우, 모형이 실제 일반적으로는 예측력이 높지 않더라도 사용한 표본 자체에 대해서만 예외적으로 높은 예측력을 갖게 될 수 있다. 따라서 모형을 측정하는 데에 사용한 표본을 통해 모형의 설명력을 설명하는 것은 신뢰성이 떨어질 수 있다.

그러므로, 유의미한 성과측정 및 비교를 위해서는 학습 표본이 아닌 표본에 대한 예측력을 논해야 한다. 가장 간단한 방법은 전체 표본을 분할하여 일부분은 모형을 추정하는 데에 사용하고, 나머지는 추정된 모형의 성능을 평가하는 데에 사용하는 것이다. 본문에서는 전체 표본을 임의분할 방식을 통해 절반으로 나누어 620,617개 중 310,308개를 학습 표본으로 사용하였고, 나머지를 평가 표본으로 분류하여 그에 대한 예측력을 비교 · 분석하였다.

이렇게 분할된 평가표본에 대한 모형의 예측성능을 비교하기 위해서는 성과측정 기준이 필요하다. 본문에서는 주택가격예측 모형에서 가장 보편적이고 직관적인 성과측정 기준인 평균 퍼센트 오차(mean absolute percent error, MAPE)와 R-squared를 사용하였다. MAPE는 모형을 통한 개별 자산의 예측치와 실제 시장가격 간의 퍼센트 오차를 평균한 것이며, 다음의 식으로 표현이 가능하다.

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{\hat{p_i}-p_i}{p_i}\right|$$

여기서 $\hat{p_i}$와 $p_i$는 각각 예측된 주택가격과 실제 주택가격을 의미한다.

R-squared는 회귀분석을 비롯한 관련 계량 분석에서 널리 사용되며, 모형을 통해 예측 가능한 표본 간 이질성이 실제의 분산에서 차지하는 비중을 의미한다. 이는 다음과 같이 표현할 수 있다.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(p_i-\hat{p_i})^2}{\sum_{i=1}^{n}(p_i-\bar{p})^2}$$

여기서 $\bar{p}$는 표본평균을 의미한다.

### 2) 입지 변수의 설정

먼저 고려할 것은 입지 변수의 설정방식이다. 주택 가격 예측에서 입지 가치의 포착은 가장 중요한 부분이다. 일반적인 헤도닉 모형에서 이는 주요 시설(도심, 교통시설, 병원, 학교, 공원 등)으로부터의 거리나 행정구역 등을 비롯한 여러 형태

로 표현되지만, 위도와 경도 자체는 변수 값으로 사용할 수 없다. 이는 헤도닉 모형이 입지 효과와 주택 가치 간의 관계를 특정한 함수로 표현하고 있기 때문이다. 그러나, 앞 장에서 알아보았듯이 랜덤 포레스트에서는 특성상 위도와 경도 자체를 통해 입지 차이에서 나타나는 가격 차이를 구분하는 방식을 택한다. 랜덤 포레스트 모형에서는 헤도닉 모형에서 사용되는 형식의 다양한 입지 변수와 위도/경도를 모두 사용할 수 있지만, 4-2절에서 논의된 바처럼 랜덤 포레스트 모형이 위도와 경도만으로 입지 가치를 충분히 포착할 수 있다면 그 이외의 변수(주요 시설로부터의 거리 또는 행정구역 등)는 모형에 포함시키지 않아도 설명력이 저하되지 않을 것이다.

〈표 2〉는 여러 가지 입지 변수 표현방식을 적용하는 경우의 예측 성능을 비교하고 있다. 먼저, 위치의 값(위도와 경도)만을 포함하는 경우에 비해 헤도닉 모형에서와 같이 주요 시설로부터의 거리와 행정동만을 포함하는 경우, 예측 성능이 오히려 저하된다는 것이다. 또한, 위치의 값, 주요 시설로부터의 거리, 그리고 행정동을 모두 포함시킨다고 하더라도 예측 성능이 위치의 값만 포함한 경우와 거의 차이가 없는 것을 확인할 수 있다. 또한, 좌표 값을 제외하고 시설로부터의 접근성과 행정동만을 포함하는 경우, 좌표 값만을 포함하는 경우에 비해 오히려 설명력이 떨어지는 것을 확인할 수 있다. 이는 앞에서 논의하였듯이 입지 효과를 포착하는 변수(여기서는 접근성, 행정구역의 구분)가 비선형적인 실제 입지 효과의 차이를 완전히 포착할 만큼 복잡하지 않기 때문이다(Adair et al., 1996; Hong et. al., 2020; Ramsey, 1969).

〈표 2〉 입지변수 표현에 따른 모형 정확도

| | MAPE | R-squared |
|---|---|---|
| 위·경도+행정동+거리 모두 포함 | 4.236 | 0.9821 |
| 위·경도만 포함 | 4.273 | 0.9818 |
| 행정동+거리 포함 | 4.514 | 0.9787 |

MAPE, mean absolute percent error.

위 정량분석 결과는 랜덤 포레스트 기반 주택 가격 모형의 경우, 위도와 경도만으로도 입지 차이에 따른 가치 차이를 잘 포착할 수 있다는 것을 보여준다. 동시에, (헤도닉 모형과 같이) 주요 시설로부터의 거리나 행정구역을 모형에 포함시키지 않아도 무방하다는 것을 의미하기도 한다.

랜덤 포레스트의 이러한 특징은 특히 모형의 실용성과 설명력 제고 측면에서 주목할 만하다. 왜냐하면 이처럼 위도와 경도만을 통해 입지 가치의 차이를 포착할 수 있다면, 그만큼 필요한 정보량이 줄어드는 것을 의미하기 때문이다. 즉, 랜덤 포레스트는 최소한의 입지 정보만으로도 복잡한 수준의 입지 가치 차이를 포착할 수 있다는 것이다. 특히, 외부효과(Nesheim, 2002)나 문화적 특성(Sheppard, 2010; Zukin, 1987)과 같이 직접 관찰하기가 어려운 효과들로 인한 입지 가치 차이에 대해서도 이러한 방법을 통하면 예측에 반영하는 것이 가능하다.

### 3) 변수의 선택 및 중요도 측정

다음으로 고려할 수 있는 것은 변수의 선택 문제이다. 선형회귀에 기반한 헤도닉 모형의 경우, 연구자가 설정한 변수가 모형 추정에 모두 사용되

므로 모형의 정확성을 높이기 위해서는 변수 설정에 대한 논의가 헤도닉 모형 자체의 추정과는 별개로 선행되어야 할 필요가 있다.

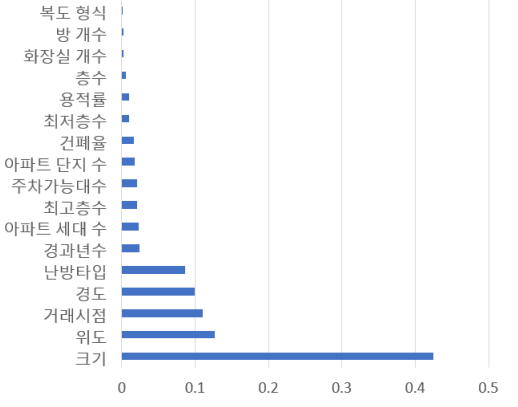
한편, 랜덤 포레스트의 경우 변수 예측에 대한 설명력(획득되는 정보량의 크기)을 기준으로 논리적 위계 구조를 형성하는 특성을 가지고 있다. 즉, 연구자가 사전에 설정한 모든 변수를 모형 추정에 사용하는 것이 아니라, 변수의 사용이 충분한 설명력을 추가하는 경우에만 계산 과정에 편입된다는 것이다. 이런 경우, 큰 정보량을 가진 변수가 원천적으로 누락되지 않은 한, 기본적으로는 어떤 변수가 사용되어야 하는 지를 연구자가 사전적으로 선별할 필요 없이 랜덤 포레스트의 알고리즘 내에서 자체적으로 선별될 수 있도록 다양한 변수를 사용해도 무방할 것이다.[5)]

다만, 랜덤 포레스트에서는 주택 가치 추정에 사용된 각 변수의 중요도를 측정할 수 있는데, 이를 통해 사후적으로 다시 변수 선택을 시도할 수 있다. 중요도는 변수를 통해 획득한 정보의 양을 통해 측정되며, 각 변수의 중요도의 총합은 1로 표준화된다.

〈표 3〉과 〈그림 4〉는 정량분석을 통해 측정된 변수의 상대적 중요도를 보여준다. 아파트 가격을 예측하는 데에 있어서, 가장 중요한 변수는 전용면적이며, 그 다음으로는 위도/경도, 거래시점, 난방타입의 중요성이 다른 변수들에 비해 확연히 높게 나타나는 것을 알 수 있다. 반면 복도형식이나 방 개수의 경우 상대적으로 예측에 기여하는 바가 낮은 것으로 나타난다. 만약 연구자가 어

〈표 3〉 변수별 중요도의 측정

| 변수 | 중요도 |
|---|---|
| 크기 | 0.4242 |
| 위도 | 0.1264 |
| 거래시점 | 0.1098 |
| 경도 | 0.099 |
| 난방타입 | 0.0865 |
| 경과 연수 | 0.0241 |
| 아파트 세대 수 | 0.0231 |
| 최고층수 | 0.0212 |
| 주차가능대수 | 0.0204 |
| 아파트 단지 수 | 0.0178 |
| 건폐율 | 0.016 |
| 최저층수 | 0.01 |
| 용적률 | 0.0095 |
| 층수 | 0.0059 |
| 화장실 개수 | 0.0025 |
| 방 개수 | 0.0018 |
| 복도 형식 | 0.0011 |




〈그림 4〉 변수별 중요도의 비교

5) 그러나 범주형 변수를 더미화하는 과정에서 변수의 개수가 불필요하게 많이 증가하는 경우, 모형의 설명력을 오히려 낮출 수도 있다. 이러한 문제는 다음 절에서 보다 자세히 논의된다.

떤 변수를 모형에서 제외함으로써 모형의 효율성을 높일 수 있는지 검사하려고 한다면, 이를 통해 상대적으로 중요도가 낮은 변수의 순서로 변수를 제외한 뒤 예측력을 비교할 수 있을 것이다.

또, 한 가지 고려할 수 있는 것은 랜덤 포레스트를 구성하는 각 의사결정나무에서 사용할 변수의 최대 개수를 설정할 수 있다는 것이다. 이는 불필요하게 많은 변수가 포함되는 경우, 오히려 과적합의 가능성을 높일 수 있기 때문이다(Segal, 2003). 앞 장에서 서술된 바와 같이, 랜덤 포레스트 기반 모형이 높은 예측력을 갖기 위해서는 과적합으로 인한 예측 성능 저하를 방지하기 위해 불필요하게 많은 변수가 포함되지 않도록 설정하는 것이 도움이 될 수 있다.

〈표 4〉는 최대 변수 개수 설정에 따른 예측 성능의 변화를 기술한 것이다. 표에서 확인할 수 있듯이, 예측 성능은 최대 변수 개수가 13개가 될 때까지 완만하게 상승하다가, 변수 개수를 더 증가시키면 오히려 성능이 저하되기 시작한다. 일반적인 선형회귀 모형에서 포함시키는 변수의 개수가 많을수록 (학습 표본에 대한) 설명력이 단조 증가하는 것과는 다른 성질을 가졌음을 확인할 수 있다.

〈표 4〉 최대 변수 개수 설정에 따른 설명력 변화

| 최대 속성 | MAPE | R-squared |
|---|---|---|
| 5 | 4.249 | 0.9822 |
| 9 | 4.188 | 0.9828 |
| 13 | 4.185 | 0.9828 |
| 17 | 4.273 | 0.9818 |

MAPE, mean absolute percent error.

#### 4) 정성 변수의 표현

다음으로, 랜덤 포레스트 모형에서 정성 변수의 효율적 적용 방법에 대해 고려해볼 수 있다. 주택 가격 예측 모형에는 다양한 변수가 사용된다. 이때 일반적으로 주택의 크기나 방 개수와 같은 정량 변수뿐 아니라, 복도 및 난방 형태나 시점과 같은 비가측 변수들도 모형에 포함된다. 헤도닉 가격 모형에서는 각 변수와 주택 가격의 관계식이 명시적으로 표현되어야 하므로, 이와 같은 비가측 변수들을 더미변수화하는 것이 일반적이다. 한편, 랜덤 포레스트 기반 모형은 각 속성값과 주택 가격의 관계식이 존재하지 않고, 속성값은 표본의 집단을 분류하는 조건식을 찾는 데에만 활용된다. 따라서 랜덤 포레스트의 경우, 더미변수를 만드는 것 이외에도 비가측 변수의 속성값을 구분 가능한 형식의 가측 변수(예를 들어, 정수의 수열)로 표현하는 방법도 생각해볼 수 있다.

이러한 표현 방식이 모형의 예측력과 효율성에 영향을 미칠 수 있는 이유는, 정성변수의 표현 방식에 따라 모형에 포함되는 변수의 수가 달라지기 때문이다. 모든 정성변수를 더미변수화하는 경우, 변수의 수는 모든 범주의 수를 합친 것만큼 늘어나게 될 것이다. 그러므로 범주의 수가 많으면 많을수록 더미화는 변수의 개수를 크게 늘리게 된다.

앞 절에서 분석한 바에 의하면, 최대 변수의 적절한 제한은 모형의 설명력을 높이는 데에 기여할 수 있다. 이와 유사한 맥락에서(자료가 유사한 정보량을 가진 경우), 불필요하게 많은 변수를 형성하는 경우, 모형의 과적합 가능성을 높이거나 연산 효율성을 낮출 가능성이 있다. 기본적으로, 랜덤 포레스트는 의사결정나무의 평균값이다. 이때 각

의사결정나무의 추정은 학습표본에서 임의 추출된 부분집합이며, 이는 해당 표본이 가지고 있는 고유한 잡음(측정 오류 등)을 포함한다. 이때 만약 변수가 지나치게 많다면, 추출된 부분 집합이 가진 잡음의 경향에 대한 설명력을 가진 변수가 나타날 가능성이 높아질 수 있으며, 이러한 변수는 (그러한 잡음을 우연히 가진 표본을 제외한) 일반 표본에 대한 설명력을 갖지 못함에도 예측 과정에 포함되어, 오히려 일반적 정보량을 가진 변수의 비중을 줄일 수 있다. 그러므로, 만약 랜덤 포레스트가 주택 시장의 정성 변수들을 정수화하는 방식으로도 그에 따른 차이를 효과적으로 포착할 수 있다면, 오히려 정성변수를 더미변수로 표현하는 경우에 비해 모형을 효율화할 수 있을 것이다.

본문에서는 일반적으로 주택가격예측에 사용되는 주요한 범주형 변수들을 정수화하는 경우와 더미변수화한 경우의 예측 성능을 비교하는 방식으로, 더 효율적인 모형 설정 방향을 찾고 있다. 본문에서 포함된 범주형 변수는 거래 시점, 주택의 복도형태, 난방방식, 아파트가 위치한 행정동이 있다. 이를 정수화하는 경우, 변수 개수는 범주형 변수의 개수와 동일하지만, 더미변수화 하는 경우 변수의 개수는 각 범주형 변수의 범주 종류만큼 더해진다. 본문의 모형에서 가장 많은 범주를 갖는 변수는 행정동으로, 400여 개의 범주수를 갖는다.

〈표 5〉는 평가 표본에 대한 랜덤 포레스트를 통한 주택가격예측 모형의 예측력이 범주형 변수를 표현하는 방식에 따라 어떻게 달라지는지를 보여준다. 확인할 수 있는 바와 같이, 예측성능지표인 MAPE와 R-squared 모두에서 더미변수화하는 경우의 예측력이 오히려 상대적으로 낮은 것으로 나타났다. 이는 상술하였듯이, 더미변수화하는 과정에서 증가한 변수의 개수가 효율적인 변수의 선택을 방해할 수 있음을 의미한다.

〈표 5〉 범주형 변수 표현 방식과 모형 정확도

| | MAPE | R-squared |
|---|---|---|
| 라벨범주 모형 | 4.236 | 0.9821 |
| 더미범주 모형 | 4.967 | 0.974 |

MAPE, mean absolute percent error.

### 5) 모형의 복잡성: 노드의 크기, 깊이

랜덤 포레스트 알고리즘은 앞 장에서 서술한 바와 같이 조건 노드의 하부에 다시 조건 노드를 생성하는 방식으로 대상을 분류한다. 이때, 조건 노드가 생성되는 조건 자체를 설정할 수 있다. 적어도 학습 표본에 대해서는 조건 노드가 가능한 한 깊이(하부 노드가 연장될 수 있는 개수), 그리고 작은 표본에 대해서도 조밀하게 생성되는 경우 예측력이 상승한다. 그러나 이런 경우 모형이 학습 표본이 가지고 있는 특성에 과적합되는 경우가 생길 수 있으며, 결과적으로 학습 외 표본에 대해서는 예측력이 오히려 저하될 가능성도 있다.

그러나, 주택가격의 예측에 있어서는 표본 간의 이질성이 높다는 것을 고려해야 할 필요성이 있다. 특히, 랜덤 포레스트는 최종 분류된 자산의 공통적 성질을 이용해 예측을 수행하므로, 최종 리프로 구성될 수 있는 표본 수에 대한 제약은 시장에서 충분히 동질성을 공유한다고 인정할 수 있는 주택 그룹의 표본 수에 대한 제약과 같다. 그러므로, 각 주택의 이질성이 본질적으로 고유한 입

지에서 비롯된다는 것(Shiller, 2015)과, 하위 시장의 존재(김경환 · 손재영, 2010; 김주영 · 우경, 2004; 손철, 2011; 정건섭 · 이상엽, 2007)를 고려하는 경우, 실제 주택가치 평가에 있어서는 상당히 조밀한 수준(경우에 따라서는 1~2개의 단지)까지 최종 분류가 이루어질 수 있어야 할 것으로 보인다. 반대로, 주택 대량평가에 있어서 노드의 분할 조건에 제약이 강하면 강할수록 실제로는 이질성을 가진 주택들이 (분할 조건에 못 미쳐서) 동질적인 주택으로 분류될 가능성이 커진다. 따라서, 랜덤 포레스트를 통해 주택 가격 예측 모형을 설계하는 경우, 노드의 생성 규칙은 주택 간 이질성을 분류하기에 충분하도록 최대한 조밀해야 할 필요성이 있다.

〈표 6〉과 〈표 7〉은 이러한 논의의 연장선 상에서 조건 노드에 대한 제약이 모형의 설명력에 어떤 영향을 미치는지를 보여주고 있다. 먼저 〈표 6〉의 횡축은 하위 노드로 분할 가능한 최소의 표본 수를 의미하는 것으로 숫자가 작을수록 더 조밀한 분류가 이루어질 수 있음을 의미한다. 종축은 분할이 불가능한 최종 노드(리프)가 될 수 있는 최소의 표본 개수에 대한 설정이다. 이때 리프는 최종적으로 동질적인 자산으로 분류된 표본의 그룹을 의미하므로, 이는 이질적인 자산으로 분류될 수 있는 최소 조건을 의미한다. 〈표 7〉은 하위 노드로 연결될 수 있는 최대 조건의 수(깊이)에 따른 예측력의 변화를 나타낸다. 모형의 깊이가 깊어질수록 더 많은 조건이 생성될 수 있으므로 모형의 더 조밀한 차이에 대해서도 포착을 하게 될 것이다.

〈표 6〉 조건 노드 생성 규칙과 모형 정확도

| Min_leaf | Min_split | | | | |
|---|---|---|---|---|---|
| | 2 | 4 | 8 | 16 | 32 |
| 1 | 4.273 | 4.263 | 4.324 | 4.574 | 5.115 |
| 3 | 4.404 | 4.404 | 4.437 | 4.658 | 5.165 |
| 5 | 4.629 | 4.629 | 4.629 | 4.770 | 5.234 |
| 10 | 5.168 | 5.168 | 5.168 | 5.168 | 5.434 |
| 20 | 6.037 | 6.037 | 6.037 | 6.037 | 6.037 |

〈표 7〉 노드 깊이와 모형 정확도

| 최대 깊이 제한 | MAPE | R-squared |
|---|---|---|
| 15 | 6.71 | 0.9703 |
| 17 | 5.46 | 0.9771 |
| 19 | 4.775 | 0.9802 |
| 21 | 4.449 | 0.9814 |
| 23 | 4.319 | 0.9817 |
| 25 | 4.276 | 0.9818 |
| 27 | 4.269 | 0.9819 |

MAPE, mean absolute percent error.

복잡성에 대한 제약이 강하면 강할수록 모형의 설명력이 하락하는 것을 표를 통해 확인할 수 있다. 〈표 6〉에서는 가장 세밀한 조건[6]에서 가장 예측력이 높으며, 〈표 7〉에서는 최대 깊이가 27이 될 때까지 예측력이 오목하게(concave) 상승하는 것을 확인할 수 있다.

이는 조건 규칙에 대한 제약이 훼손시키는 설명력이 잠재적 과적합으로 인한 예측력의 감소를 상회함을 의미할 수 있다. 제약 강화에 따른 설명

6) 최소 분할 조건이 2라는 것과 최종 노드가 될 수 있는 최소 표본 개수가 1이라는 것은 노드 분할에 아무 조건이 없다는 것과 같은 의미이다.

 https://doi.org/10.30902/jrea.2021.7.1.1

력의 감소가 의미하는 것은 분할에 필요한 요구조건이 강해지면서 실제로는 이질성을 가진 주택들이 동질적인 주택으로 분류되었다는 것이다. 동시에, 이는 모형이 실제 하위시장의 복잡성과 특이성을 포착할 만큼 조밀하게 확장되지 못한 데에서 생긴 설명력의 저하라고 해석할 수도 있다.

### 6) 시점의 분할

앞의 4-3절에서 논의한 바에 따르면, 랜덤 포레스트는 거시경제적 변화가 나타나는 각 시점별로 속성과 가격의 관계를 다르게 구조화할 수 있기 때문에, 연구자가 사전에 자료의 동질성을 해치지 않는 범위의 시점을 설정하지 않아도 모형의 설명력이 높게 유지될 것임을 예상할 수 있다.

이는 헤도닉 모형을 통한 대량평가와 상당히 다른 특징이다. 헤도닉 모형에서도 시점에 따른 효과를 모형에 더미변수 등의 형태로 포함시킬 수는 있지만, 기본적으로 시점마다 달라질 수 있는 변수 간의 관계나 입지 가치의 불균등한 변화 등을 고려하면, 모형 추정에 포함된 시점의 구간이 길어지면 길어질수록 설명력의 훼손이 발생할 것임을 알 수 있다. 예를 들어, 김우성 외(2019)는 서울 강남 아파트에 대한 분석을 통해 각 변수가 가격에 미치는 영향이 시간 변화에 따라 변해간다는 것을 보여주었으며, 이는 헤도닉 모형의 추정에 사용되는 자료의 시점상 동질성이 연구자에 의해 사전적으로 설정되어야 한다는 것을 의미한다. 그러므로, 본문에서 사용된 것과 같이 자료가 장기(약 10년)에 걸쳐 있는 경우, 헤도닉 모형의 경우 연구자가 사전에 시점을 분할해서 별개의 모형으로 추정하는 것이 일반적으로 모형의 설명력을 유지할 수 있는 방식이다.[7)]

〈표 8〉은 본문에서 사용한 자료(2009.7~2019.12)를 4개의 시점으로 분할하여 각각 랜덤 포레스트를 통해 추정한 결과이다. 위와 동일하게 각 시점별로 학습 표본과 평가 표본을 50%씩 임의 분할하였으며, 표의 정확도는 평가 표본에 대한 예측력을 의미한다. 표에서 알 수 있듯이, 랜덤 포레스트의 경우 전 기간의 표본을 동시에 사용하더라도 설명력의 훼손이 일어나기보다 오히려 예측력이 소폭 상승하는 것을 확인할 수 있다. 이는 상술된 바와 같이 랜덤 포레스트의 경우, 시점 변화에 따른 효과를 자체적으로 통제할 수 있기 때문에 설명력 훼손이 일어나지 않고, 오히려 전 기간에 해당하는 표본을 사용하는 경우 분할된 시점별 표본을 사용하는 경우에 비해서 더 많은

〈표 8〉 헤도닉 모형과 랜덤 포레스트

| | MAPE | R-squared |
|---|---|---|
| 기간 1 (2009.7~2012.2) | 5.578 | 0.962 |
| 기간 2 (2012.2~2014.9) | 4.443 | 0.973 |
| 기간 3 (2014.9~2017.4) | 3.98 | 0.981 |
| 기간 4 (2017.4~2019.12) | 4.598 | 0.979 |
| 전 기간 (2009.7~2019.12) | 4.236 | 0.982 |

MAPE, mean absolute percent error.

7) 예를 들어, IAAO에서 제시하고 있는 대량평가 표준에서는 가격 변동성에 따른 예측치의 왜곡을 우려하여 모형 추정에 사용되는 자료를 3년 이내로 제한하라고 권고하고 있다.

표본을 통해 모형을 학습할 수 있기 때문에 예측 성능이 향상되는 것을 발견할 수 있다.

### 7) 헤도닉 모형과의 비교

마지막으로, 본문은 랜덤 포레스트 모형와 헤도닉 가격 모형을 통한 주택가격 평가의 예측력을 비교하였다. 앞 장에서 서술한 바와 같이, 랜덤 포레스트 모형은 1. 하위 시장의 존재로 인한 비선형성을 포착하기에 특화되어 있으며, 2. 주택의 위치(위도와 경도)의 차이를 통해 입지 가치의 차이를 포괄적으로 예측에 반영할 수 있다. 그러므로 1. 주택 시장에 하부 시장이 복잡하게 존재하면 할수록, 2. 입지 가치의 결정요소가 복잡하고 비선형적일수록 헤도닉 모형에 비해 랜덤 포레스트 기법의 예측력이 더 높게 나타날 가능성이 크다. 동시에, 랜덤 포레스트의 예측력이 헤도닉 모형에 비해 높다는 것은 그러한 복잡성(하부시장 또는 관측 어려운 입지효과의 존재)이 주택 가격 결정 구조에 일으키는 비선형성이 그만큼 크다는 것을 의미한다.

〈표 9〉는 전통적 헤도닉 모형과 랜덤 포레스트 모형의 학습 외 표본에 대한 예측력을 비교한 것이다. 헤도닉 모형의 경우, 랜덤 포레스트 모형과 동일한 구조적 특성들을 적용하였으며, 입지 가치를 포착하기 위해서 주요 시설들로부터의 거리 뿐 아니라 행정동 더미까지 동시에 고려하였다. 통제 변수 간의 내생성이 발생하는 경우라도 종속변수에 대한 예측력은 손상되지 않다는 것을 고려하면(Greene, 2003), 이처럼 많은 변수를 고려하는 것이 헤도닉 모형의 예측 성능을 극대화하는 방법이 될 수 있을 것으로 보인다.

〈표 9〉 헤도닉 모형과 랜덤 포레스트

| | MAPE | R-squared |
|---|---|---|
| 랜덤 포레스트 | 4.1857 | 0.9828 |
| 헤도닉 모형 | 11.51 | 0.8964 |

MAPE, mean absolute percent error.

입지에 따른 평균적 소득 수준·편의성·학군의 차이 등 주요한 요소들이 행정동의 단위로 상당 부분 포착될 수 있다는 점을 고려할 때, 두 모형의 예측력 차이는 상술된 비선형성의 포착 여부에서 비롯된다는 것을 추측할 수 있다. 동시에 이는 랜덤 포레스트가 기술적으로 적절하게 포착할 수 있다는 실증적 근거가 될 수 있다.

헤도닉 모형의 경우, 학습 외 표본에 대한 평균 오차는 약 11.5%, R-squared는 0.896이다. 기본적으로 헤도닉 모형이 선형회귀에 기반했다는 것과, 본문 서울 전 지역을 대상으로 하는 대량의 표본(320,308개)을 사용하고 있음을 감안하면 상당히 높은 설명력을 가졌다는 것을 알 수 있다. 그럼에도 불구하고, 이는 랜덤 포레스트의 예측력과 큰 차이가 있는 것으로 나타난다. 랜덤 포레스트의 경우, 입지 가치의 차이를 포착하기 위해 위도와 경도 변수만 고려하였으며, 여타 범주형 변수는 모두 라벨 형식으로 표현하였다. 또한, 최대변수의 크기는 13개로 제한되었다. 이 경우, 랜덤 포레스트를 통한 예측의 평균 오차는 4.18%에 불과하며, R-squared의 경우 0.98로 거의 대부분의 분산이 랜덤 포레스트 모형에 의해 설명 가능한 것으로 확인되기 때문이다.

## Ⅵ. 결론

본문은 랜덤 포레스트 알고리즘을 통해 주택 가격 결정구조의 복잡성을 보다 효율성으로 포착할 수 있음을 논의하였으며, 특히 랜덤 포레스트 모형이 하부 시장으로 인한 비선형성과 입지효과의 포착에 효과적임을 분석하였다. 또한, 본문은 이를 바탕으로 랜덤 포레스트 모형 내에서 주택 시장 변수를 처리하는 방식에 모형의 예측력이 어떻게 변화하는지를 제시하고, 전통적인 선형 헤도닉 모형과의 예측을 비교하였다. 본문은 2009년부터 2019년 사이 서울에서 거래된 아파트 620,217건을 표본으로 사용하였다. 이는 헤도닉 모형 분석을 포함한 기존 연구들에서 사용되는 일반적인 표본의 수와 범위에 비해서 상당히 큰 것으로, 표본 크기에 따른 모형의 예측력 상승을 기대할 수 있을 뿐 아니라, 본문에서 분석된 결과의 일반성을 뒷받침할 수 있을 것이다.

그럼에도 불구하고, 본문에서 제시된 평가모형은 주택 가격의 예측 시 사용된 정보의 시점상 선후 관계를 자세히 통제하고 있지 않다는 한계점을 가지고 있다. 주택 평가의 활용 시에는 현재의 주택 가격뿐 아니라, 과거 특정 시점의 자산 가치를 추정할 필요성이 존재하는데, 이때 최대한의 표본 수를 이용하여 예측력을 높이기 위해서는 이와 같은 방식이 사용될 수 있다. 그러나 이러한 모형을 현재 또는 근미래의 주택 가격을 예측하는 데에 활용하는 경우에는, 이용할 수 있는 정보가 현재까지의 정보로 국한되므로 모형의 예측력이 본문에서 제시된 것에 비해 다소 낮을 것이다. 이러한 한계에 대한 기술적 극복 역시 추후 흥미로운 연구 과제가 될 것으로 보인다.

또, 한 가지 본문에서 소개된 랜덤 포레스트 기법의 한계점은 대량평가 등 가격 예측 이외의 분석으로의 그 확장성이 제한되어 있다는 것이다. 이는 기계학습 모형의 특성상, 변수 간 관계에 대한 표현이 직관적이지 않을 뿐더러 개별 변수의 효과를 명료히 관찰할 수 없기 때문에 나타나는 구조적 한계점에 기인한다. 즉, 랜덤 포레스트라는 방법론 자체가 일반적인 헤도닉 모형과 같이 세밀하고 직관적인 사회분석보다는 (그러한 분석력을 포기하더라도) 결과의 예측성능을 높이는 데에 최적화되어 있다는 것이다.


## ORCID

홍정의 https://orcid.org/0000-0002-0591-3285


## 참고문헌


1. 구본창 · 송현영, 2001,「아파트 특성에 따른 가격결정모형 연구: 분당신도시를 대상으로」,『주택포럼』, 16(2): 136-143.
2. 김경환 · 손재영, 2010,『부동산 경제학』, 서울:건국대학교 출판부.
3. 김덕중, 2002,「헤도닉 모형을 이용한 아파트 가격결정요인과 가치추정에 관한 연구」, 건국대학교 석사학위논문.
4. 김민성 · 박세운, 2014,「지하철 접근성이 아파트 가격에 미치는 영향에 관한 연구」,『한국경영학회 통합학술발표논문집』, 2912-2931.
5. 김종수 · 이성근, 2012,「헤도닉가격모형과 서포트

벡터 회귀분석모형을 이용한 공업용 부동산의 가격 추정」, 『감정평가학논집』, 11(1): 71-89.

6. 김주영 · 우경, 2004, 「수도권 주택하위시장 분석에 관한 연구」, 『국토연구』, 41: 101-111.
7. 김우성 · 이시온 · 장현수 · 김재완 · 홍정의, 2019, 「헤도닉 가격 모형을 통한 주거 선호의 구조 변화 분석: 2006~2017년 강남 지역 아파트를 중심으로」, 『부동산학보』, 76: 137-150.
8. 김운정, 2004, 「아파트 평면이 가격에 미치는 영향: 서울 아파트 밀집지역을 중심으로」, 건국대학교 석사 학위논문.
9. 김진유 · 이창무, 2005, 「어메니티요소가 주택가격에 미치는 영향력의 시계열적 변화」, 『국토계획』, 40(1): 59-74.
10. 김천일, 2018, 「경과년수와 용적률의 상호 작용을 고려한 아파트 가격 형성 분석」, 『부동산 분석』, 4(1): 1-14.
11. 김태훈 · 홍한국, 2004, 「회귀모형과 신경망모형을 이용한 아파트 가격 모형에 관한 연구」, 『국토연구』, 43: 183-200.
12. 배성완 · 유정석, 2018, 「머신 러닝 방법과 시계열 분석 모형을 이용한 부동산 가격지수 예측」, 『주택연구』, 26(1): 107-133.
13. 손철, 2011, 「수도권의 공간적 주택하위시장 식별에 대한 연구」, 『국토연구』, 70: 151-166.
14. 손아남 · 정경수, 2014, 「아파트 주변 공원녹지의 속성 가치, 『상경연구』, 39(1): 1-21.
15. 원두환 · 김형건, 2008, 「난방 방식에 따른 아파트 가격 변화 분석」, 『에너지경제연구』, 7(2): 75-101.
16. 윤정중, 2001, 「도시경관의 조망특성이 주택가격에 미치는 영향」, 연세대학교 박사학위논문.
17. 윤채규, 2003, 「주관적 평가요인을 고려한 공동주택의 가격 결정모형 개발: 분당 · 일산 신도시 공동주택을 중심으로」, 연세대학교 박사학위논문.
18. 이강 · 최근희, 2016, 「헤도닉 가격모형을 활용한 주택가격 결정요인에 관한 연구」, 『한국도시행정학회 학술발표대회 논문집』, 317-333.
19. 이창로 · 박기호, 2016, 「단독주택가격 추정을 위한 기계학습 모형의 응용」, 『대한지리학회지』, 51(2): 219-233.
20. 정건섭 · 이상엽, 2007, 「주택하위시장 구분방법과 정책적 시사점」, 『한국정책분석평가학보』, 17(1): 193-216.
21. 정수연 · 김태훈, 2007, 「헤도닉모형을 이용한 아파트 층별효용비율에 관한 연구: 서울지역을 대상으로」, 『부동산연구』, 17(1): 27-48.
22. 조주현, 1998, 「주택밀도가 주택가격에 미치는 영향에 관한 연구」, 『사회과학논총』, 건국대학교 사회과학연구소.
23. 최윤아 · 송병하, 2006, 「공동주택가격에 영향을 미치는 주거환경 요소의 중요도 평가에 관한 연구」, 『대한건축학회논문집 계획계』, 22(11): 115-124.
24. 하유정 · 이현석, 2020, 「교육환경이 아파트 가격에 미치는 영향: 부산시를 중심으로-」, 『부동산 도시연구』, 13(1): 47-61.
25. 홍정의, 2020, 「기계학습 알고리즘을 이용한 주택가격감정 시스템의 구축 및 평가: XGBoost, Light GBM, CatBoost 알고리즘에 기반하여」, 『주택금융연구』, 4: 33-64.
26. Adair, A. S., J. N. Berry, and W. S. McGreal, 1996, "Hedonic modelling, housing submarkets and residential valuation," *Journal of Property Research*, 13: 67-83.
27. Antipov, E. A. and E. B. Pokryshevskaya, 2012, "Mass appraisal of residential apartments: An application of random forest for valuation and a CART-based approach for model diagnostics," *Expert Systems with Applications*, 39: 1772-1778.
28. Alonso, W., 1964, *Location and Land Use. Toward a General Theory of Land Rent.*

Cambridge, MA: Harvard University Press.

29. Čeh, M., M. Kilibarda, A. Lisec, and B. Bajat, 2018, "Estimating the performance of random forest versus multiple regression for predicting prices of the apartments," *ISPRS International Journal of Geo-Information*, 7: 168.
30. Fan, G. Z., S. E. Ong, and H. C. Koh, 2006, "Determinants of house price: A decision tree approach," *Urban Studies*, 43: 2301-2315.
31. Glaeser, E. L., 2019, *Agglomeration Economics*. Chicago, IL: The University of Chicago Press.
32. Gu, J., M. Zhu, and L. Jiang, 2011, "Housing price forecasting based on genetic algorithm and support vector machine," *Expert Systems with Applications*, 38: 3383-3386.
33. Greene, W. H., 2003, *Econometric Analysis*, 5th ed. Harlow, UK: Pearson Education.
34. Heyman, A. V., S. Law, and M. Berghauser Pont, 2019, "How is location measured in housing valuation? A systematic review of accessibility specifications in hedonic price models," *Urban Science*, 3: 3.
35. Hong, J., H. Choi, and W. S. Kim, 2020, "A house price valuation based on the random forest approach: The mass appraisal of residential property in South Korea," *International Journal of Strategic Property Management*, 24(3): 140-152.
36. Limsombunchai, V., 2004, "House price prediction: Hedonic price model vs. artificial neural network," In *New Zealand Agricultural and Resource Economics Society Conference*, Blenheim, New Zealand, 25-26.
37. Malpezzi, S., 2002, "Hedonic pricing models: A selective and applied review," *Housing Economics and Public Policy*, 67-89.
38. McCluskey, W. and S. Anand, 1999, "The application of intelligent hybrid techniques for the mass appraisal of residential properties," *Journal of Property Investment & Finance*, 17: 218-239.
39. Mu, J., F. Wu, and A. Zhang, 2014, "Housing value forecasting based on machine learning methods," *Abstract and Applied Analysis*, 2014: 648047.
40. Nesheim, L., 2002, "Equilibrium sorting of heterogeneous consumers across locations: Theory and empirical implications," *Cemmap Working Paper*, No. CWP08/02.
41. Osland, L. and I. Thorsen, 2008, "Effects on housing prices of urban attraction and labor-market accessibility," *Environment and Planning A: Economy and Space*, 40: 2490-2509.
42. Ramsey, J. B., 1969, "Tests for specification errors in classical linear least-squares regression analysis," *Journal of the Royal Statistical Society: Series B (Methodological)*, 31: 350-371.
43. Rosen, S., 1974, "Hedonic prices and implicit markets: Product differentiation in pure competition," *Journal of Political Economy*, 82: 34-55.
44. Segal, M. R., 2003, *Machine Learning Benchmarks and Random Forest Regression*, Dordrecht, Netherlands: Kluwer Academic.
45. Selim, H., 2009, "Determinants of house prices in Turkey: Hedonic regression versus artificial neural network," *Expert Systems with Applications*, 36: 2843-2852.
46. Sheppard, S., 1999, "Hedonic analysis of housing markets," *Handbook of Regional and Urban Economics*, 3: 1595-1635.
47. _____, 2010, *Measuring the Impact of Culture Using Hedonic Analysis*, Williamstown, MA: Center for Creative Community Development.
48. Shiller, R. J., 2015, *Irrational Exuberance:*

*Revised and Expanded*, 3rd ed. Princeton, NJ: Princeton University Press.

49. Tabuchi, T., 1998, "Urban agglomeration and dispersion: A synthesis of Alonso and Krugman," *Journal of Urban Economics*, 44: 333-351.
50. Wang, D. and V. J. Li, 2019, "Mass appraisal models of real estate in the 21st century: A systematic literature review," *Sustainability*, 11: 7006.
51. Watkins, C. A., 2001, "The definition and identification of housing submarkets," *Environment and Planning A: Economy and Space*, 33: 2235-2253.
52. Wheaton, W. C. and M. J. Lewis, 2002, "Urban wages and labor market agglomeration," *Journal of Urban Economics*, 51: 542-562.
53. Woods, E. and E. Kyral, 1997, *Ovum Evaluates Data Mining*, London, UK: London Ovum.
54. Zhou, G., Y. Ji, X. Chen, and F. Zhang, 2018, "Artificial neural networks and the mass appraisal of real estate," *International Journal of Online and Biomedical Engineering*, 14: 180-187.
55. Zurada, J., A. Levitan, and J. Guan, 2011, "A comparison of regression and artificial intelligence methods in a mass appraisal context," *Journal of Real Estate Research*, 33: 349-388.
56. Zukin, S., 1987, "Gentrification: Culture and capital in the urban core," *Annual Review of Sociology*, 13: 129-147.

**논문접수일**: 2021년 3월 11일
**심사(수정)일**: 2021년 4월 12일
**게재확정일**: 2021년 4월 23일

**국문초록**

본문은 최근 주목받는 기계학습 기법인 랜덤 포레스트 모형이 주택시장의 복잡성을 포착하는 데에 어떻게 활용될 수 있는지를 조명하고, 그것을 바탕으로 효율적 모형 설계를 위한 다양한 정량분석을 시도하였다. 분석 결과는 다음과 같이 정리된다. 첫째, 랜덤 포레스트 모형은 하부시장이나 입지효과의 비선형성으로 인한 복잡한 가격 차이를 포착하는 데에 유용하게 사용될 수 있다. 선형 헤도닉 모형과의 예측력 비교에서, 랜덤 포레스트 기반 모형의 평균 오차 백분율은 약 4%로 헤도닉 모형(약 11%)에 비해 크게 낮은 것으로 나타났다. 둘째, 랜덤 포레스트는 입지효과를 나타내는 대리변수를 사용할 필요 없이 위치 정보만으로도 입지가치의 차이를 예측에 반영할 수 있다. 셋째, 정성변수의 더미화는 연산에 포함되는 변수의 수를 크게 증가시키므로 지수화에 비해 오히려 모형의 설명력을 낮출 가능성이 높다. 넷째, 주택 시장의 복잡성을 예측에 충분히 반영하기 위해 랜덤 포레스트의 조건 분할 역시 가능한 한 세밀하게 설정되는 것이 유리하다. 즉, 모형의 복잡성이 상승시키는 설명력이 과적합으로 인한 설명력 훼손보다 큰 경향이 있다. 다섯째, 랜덤 포레스트 모형의 추정 시에는 연구자가 사전에 표본을 기간 동질성을 갖도록 분할 할 필요가 없다.

주제어 : 기계학습, 랜덤 포레스트, 주택 평가 모형, 하부시장, 입지효과