pISSN : 2465-9754 eISSN : 2508-1292 https://doi.org/10.30902/jrea.2025.11.2.21

Journal of Real Estate Analysis

http://www.ejrea.org *July 2025, Vol.11, No.2, pp. 21~40*

XGBoost 기반 부동산 자동가치산정모형 (Automated Valuation Model)의 실증 분석
- 은평구 다세대주택 실거래가를 중심으로 -

Empirical Analysis of XGBoost-based Real Estate Automated Valuation Model
- Actual Transaction Prices of Multi-family Housing in Eunpyeong-gu -

이선구*
Seon Ku Lee

▌Abstract ▐

This study empirically examines the predictive performance of an automated valuation model (AVM) with the XGBoost regression algorithm based on actual transaction data on multi-family housing in Eunpyeong-gu, Seoul. Unlike prior studies that typically apply Min-Max normalization to address variable-scale discrepancies, this study maintains the original transaction price units without normalization, thereby enhancing the interpretability and practical utility of the results in real-world applications. The dataset consists of 1,839 real transactions recorded between November 2023 and October 2024, of which 1,272 were used for model training and 567 for testing. The predictive model integrates a wide range of variables including physical attributes, locational and environmental factors, as well as market and macroeconomic indicators. The empirical results demonstrated a test score (normalized root mean squared error) of 0.136511, coefficient of determination (R^2) of 0.747, and mean absolute percentage error of 13.6%. These results suggest that the model can approximate the actual market prices with an average accuracy of approximately 86.4%. Given that the average transaction price during the test period was approximately KRW 291.63 million, the model's mean prediction error corresponds to approximately KRW 39.66 million. The results confirm that the XGBoost model effectively captures the nonlinear and heterogeneous nature of real estate transaction data. Generating predictions in raw monetary units rather than in normalized values provides a more intuitive and practically interpretable AVM structure and offers a viable alternative to conventional normalization-based frameworks. Furthermore, this study empirically demonstrates the applicability

* 동서울대학교 도시계획·부동산학과 겸임교수, 세종디엑스 디지털에셋사업팀 리더 | Adjunct Professor, Department of Urban Planning & Real Estate, Dong Seoul University; Leader, Digital Asset Business Team, SejongDX | leeseonku@gmail.com |

and explanatory power of machine learning models in real estate valuation, contributing both academically and practically to the development of more usable, sophisticated AVM systems.

Keywords: Automated valuation model (AVM), XGBoost, Real estate price prediction, Machine learning, Multi-family housing

Ⅰ. 서론

부동산 시장은 전통적으로 전문가의 경험과 비교사례 접근법에 기반한 가격 산정이 이루어져 왔으나, 이러한 방식은 주관성이 개입될 여지가 크고, 대규모 데이터 기반의 신속한 가치평가에는 한계가 있다. 특히 정보 비대칭성과 자산 간 이질성은 예측의 정확도를 저해하는 구조적 제약으로 작용해 왔다. 이에 따라 데이터 기반 자동화 가치평가 모델인 AVM(automated valuation model)의 도입 필요성이 대두되고 있으며, 금융기관 담보평가, 리스크 분석, 프롭테크 시장 예측 등에서 활용 가능성이 확대되고 있다.

AVM은 실거래 데이터를 기반으로 부동산 가치를 자동 산정하는 모델로, 시간·비용 측면에서 효율성이 높고, 다양한 실무 영역에 적합한 도구로 평가된다. 다만, AVM의 성능은 지역, 자산 유형, 데이터 질, 알고리즘에 따라 민감하게 달라지므로, 특정 유형을 대상으로 한 실증 검증이 요구된다.

본 연구는 서울시 은평구 다세대주택 실거래 1,839건을 기반으로, XGBoost(eXtreme gradient boosting) 알고리즘을 적용한 AVM을 구축하고 그 예측 성능과 실무 활용 가능성을 검토한다.

연구 목적은 다음과 같다. 첫째, XGBoost 기반 AVM의 예측 정확도와 실무 적용 가능성을 평가하고, 둘째, 부동산 가격 결정에 유의한 주요 변수들을 식별하며, 셋째, 자동평가 시스템이 정보 비대칭성과 가격 투명성 문제 해결에 기여할 수 있는지를 검토하는 데 있다.

본 연구의 범위는 부동산 AVM의 예측 성능을 실증적으로 검증하기 위해 공간적, 시간적, 자산적, 자료적 측면에서 진행하였다. 첫째, 공간적 범위는 서울특별시 은평구로 한정하였다. 은평구는 서울 서북권에 위치하며, 중저가 주거시장을 대표하는 지역으로서 다세대 및 연립주택의 비중이 높고, 주거유형의 다양성과 안정적인 거래량이 확보된 지역이다. 실거래가 자료를 분석한 결과, 강서구(raw data 2,550건)와 은평구(raw data 2,446건)가 서울시 25개 자치구 중 가장 높은 거래건수를 기록하였으며, 그 외 광진구(raw data 1,595건), 송파구(raw data 1,521건), 강북구(raw data 1,430건), 중랑구(raw data 1,416건) 등도 상대적으로 활발한 거래를 보였다. 그러나 강서구는 최근 수년간 전세사기 피해가 집중된 지역으로, 실증분석 대상지로서의 적합성이 낮다고 판단하였다. 이에 비해 은평구는 실수요 기반의 안정적인 거래 구조를 유지하고 있으며, 재개발·재건축 등 도시정비사업의 영향도 제한적으로 나타나 단기 외생변수에 의한 가격 왜곡 가능성이 낮다. 따라서 은평구는 지역적 이질성과 주택 유형별 특성을 반영한 부동산 가치 변동을 실

 https://doi.org/10.30902/jrea.2025.11.2.21

증적으로 관찰하기에 적합한 분석대상지로 판단된다. 둘째, 시간적 범위는 2023년 11월부터 2024년 10월까지 1년으로 설정하였다. 이는 계절성, 경기 변동성, 거시경제 환경 변화를 모두 반영할 수 있는 기간으로, 데이터의 시계열적 일관성과 현실 반영성을 동시에 확보하고자 하였다. 셋째, 자산 범위는 소규모 개별성 및 물리적 다양성이 강한 다세대 및 연립주택으로 한정하였다. 이들 자산군은 아파트와 달리 구조적 이질성이 커 AVM의 예측력을 검증하는 데 이론적 도전성과 실무적 의의가 높다. 넷째, 자료는 국토교통부 실거래가 공개시스템을 통해 수집한 공식 거래 사례로 구성하였으며, 통계적 분석과 머신러닝 기반 모형 학습에 적당한 규모를 갖춘 것으로 판단된다.

이와 같은 연구 범위 설정을 통해 본 연구는 딥러닝 기반 AVM의 실질적 효용성과 적용 가능성을 정밀하게 검토하고자 한다.

연구의 방법에 있어서는 부동산 AVM의 구축 및 예측 성능 검증을 위하여 다음과 같은 절차를 따랐다. 첫째, 자료는 서울시 은평구 연립 및 다세대주택의 실거래가를 대상으로 하였으며, 국토교통부 실거래가 공개시스템에서 수집한 공신력 있는 데이터를 사용하였다. 거래의 신뢰성을 확보하기 위하여 직거래 및 계약해제 사례는 모두 제외하였고, 분석 변수는 부동산 특성(전용면적, 대지지분 등), 입지 요인(지하철 접근성, 성범죄자 거리), 시장 환경(소비심리지수, 가격지수), 거시경제 변수(시장금리) 등으로 구성하였다. 둘째, 변수 구성은 부동산 가치결정 이론에 근거하였으며, 일부 변수에는 로그 변환을 적용하여 데이터의 정규성을 제고하였다. 셋째, 분석모형은 XGBoost regressor를 기반으로 하였으며, 전체 데이터는 학습용(70%)과 검증용(30%)으로 분할하고, grid search 및 5-fold cross validation을 통해 최적의 하이퍼파라미터를 도출하였다. 과적합 방지를 위해 Early Stopping 기법을 병행 적용하였다. 넷째, 예측 성능 평가지표는 MAPE(mean absolute percentage error)와 R^2(결정계수)를 사용하였으며, 기존 연구와 달리 Min-Max(min-max normalization) 정규화를 적용하지 않고 실거래 금액의 원단위를 그대로 사용함으로써 실제값과 예측값 간의 직관적 비교가 가능하도록 하였다. 다섯째, 변수 중요도 분석을 통해 각 독립변수의 예측 기여도를 정량적으로 도출하였다. 이와 같은 분석 절차를 통해 본 연구는 XGBoost 기반 AVM의 실무적 활용 가능성과 한계를 평가하고, 향후 부동산 데이터 기반 가치평가 체계 고도화에 기여하고자 한다.

II. 이론 및 선행연구 고찰

1. 이론적 배경

1) Automated Valuation Model

전통적인 감정평가 방식은 비교사례 접근법, 원가 접근법, 수익환원법 등에 기반하나, 평가자의 주관성과 경험에 따라 결과가 달라질 수 있다는 구조적 한계를 내포하고 있다. 특히 다세대·연립주택처럼 비정형성이 높은 자산에 대해서는 유사사례의 선정과 보정계수 산정이 정량적 기준 없이 이루어지는 경우가 많아, 평가 결과의 재현성

과 표준화가 어렵다. 또한 현장 실사와 사례 수집 등 복잡한 절차는 시간과 비용이 많이 소요되어 대량 데이터 기반의 실시간 분석이 필요한 프롭테크·금융 실무와의 정합성에도 한계가 있다. 이에 반해, 기계학습 기반의 AVM은 객관적 데이터 기반의 자동화된 예측이 가능하다는 점에서 기존 감정평가 방식의 주관성과 비효율성을 실질적으로 보완할 수 있는 유의미한 대안으로 평가된다.

또한, 2000년대 이후 실거래가 공개, 건축물대장의 디지털화, 공간정보 기술의 발달 등으로 부동산 정보의 전산화와 정형화가 이루어지면서, 계량적 가치평가의 가능성이 확대되었다. 이와 같은 데이터 환경은 통계적 추론 및 기계학습 기법의 부동산 분야 도입을 촉진하였다.

AVM은 일정한 알고리즘과 입력 절차를 통해 대규모 거래 및 자산 속성 데이터를 학습하고, 반복적·체계적으로 부동산 가치를 산정하는 모델이다. AVM은 평가의 효율성과 표준화를 실현함으로써 감정평가의 보조 수단으로서, 프롭테크 산업과 금융·공공부문 전반에 걸쳐 활용 가능성이 확대되고 있다.

2) XGBoost 알고리즘

XGBoost는 Friedman(2001)이 제안한 GBDT (gradient boosted decision tree) 알고리즘의 개선 버전으로, 앙상블 학습(ensemble learning)의 일종인 부스팅(boosting) 기법을 확장하여 높은 예측 정확도, 계산 효율성, 정규화 기반 일반화 능력을 동시에 확보한 고성능 회귀·분류 알고리즘이다. 해당 알고리즘은 수천만 개의 관측치와 고차원의 피처 집합을 갖는 비정형 데이터셋에 대해서도 높은 학습 안정성과 확장성을 제공하는 것으로 알려져 있다. XGBoost의 목적함수는 예측 오차를 최소화하는 손실함수(loss function)와 모델 복잡도에 대한 정규화항(regularization term)의 합으로 구성된다. 이는 다음 (식 1)과 같이 정의된다.

$$L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{k} \Omega(f_k) \quad \text{(식 1)}$$

여기서 l은 예측값 $\hat{y}_i$와 실제값 y_i 간의 손실을 측정하는 함수(보통 제곱오차 혹은 로그우도), f_k는 k-번째 단계의 결정트리이며, $\Omega(f_k)$는 트리 복잡도에 대한 패널티 함수로 정의된다. 특히 $\Omega(f_k)$는 다음 (식 2)와 같이 구성된다.

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \quad \text{(식 2)}$$

여기서 T는 리프 노드의 개수, ω_j는 각 리프 노드의 가중치, γ는 노드 수에 대한 정규화 계수, λ는 가중치 크기에 대한 L_2 규제항이다. 이러한 정규화 항은 모델의 복잡도를 제어함으로써 과적합(overfitting)을 방지하는 역할을 수행한다.

XGBoost는 목적함수를 테일러 급수(Taylor expansion)를 이용하여 2차 근사한 후, 반복적인 잔차 학습(residual fitting) 과정을 통해 각 반복(iteration)에서의 최적 트리를 구축한다.

t-번째 반복에서의 손실함수는 다음과 같이 근사된다.

$$L^{(t)} \approx \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \quad \text{(식 3)}$$

여기서 $g_i = \frac{\partial l(y_i, \hat{y_i}^{(t-1)})}{\partial \hat{y_i}^{(t-1)}}$는 1차 도함수(그래디언트), $h_i = \frac{\partial^2 l(y_i, \hat{y_i}^{(t-1)})}{\partial \hat{y_i}^{(t-1)^2}}$는 2차 도함수(헤시안)로서, 이차 근사를 통해 최적의 리프 노드 분할 및 가중치 업데이트가 이루어진다. 이러한 구조는 기존 GBDT 대비 학습 속도와 안정성을 획기적으로 향상시킨다. 각 반복 단계에서 트리 구조는 이차 도함수 기반의 최적 리프 분할 기준을 바탕으로 생성된다. 리프 노드 j의 이득(gain)은 다음과 같이 정의된다.

$$Gain = \frac{1}{2}[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}] - \gamma \quad \text{(식 4)}$$

여기서 G_L, G_R은 좌우 리프의 그래디언트 누적합, H_L, H_R는 좌우 리프의 헤시안 누적합이다. Gain 값이 양수인 경우 해당 분할은 모델 성능 향상에 기여하며, 이득이 가장 큰 분할 기준을 선택하여 트리 구조를 확장한다. 이를 통해 모델은 각 변수의 국지적 중요도(local importance)를 학습하게 되며, 변수 간의 비선형적 상호작용과 상위 변수 조합(high-order interactions)까지도 효과적으로 포착할 수 있다. XGBoost는 결측치(missing values)에 대해 자동 처리를 수행하며, 관측값이 누락된 경우 각 노드에서 최적 경로를 역추적하여 누락 분기를 자동 결정한다. 또한 일정 epoch 이상 성능 향상이 없는 경우 학습을 자동 중단하는 조기 종료(early stopping) 기법을 지원하여 과적합을 억제한다.

모델이 학습한 변수 중요도는 gain, cover, frequency 등 다양한 기준으로 시각화가 가능하며, 특히 본 연구에서는 이를 수치화하였다. 이러한 특성은 선형회귀모형이나 고전적 통계모형이 갖는 구조적 제한성과 모형 미지정 오류(model misspecification)의 가능성을 증가시키며, 일반화 성능의 저하를 초래한다. 반면 XGBoost는 위와 같은 비정형 특성을 구조적으로 포용할 수 있는 알고리즘적 유연성과 수학적 정교함을 동시에 갖추고 있으며, 학습 속도 및 성능 측면에서도 기존의 GBDT, 랜덤포레스트(random forest), SVM(support vector machine) 등의 기법을 상회하는 것으로 평가된다.

특히, 본 연구의 대상인 다세대주택은 거래량, 입지조건, 건축특성 등에서 표준화된 아파트와는 달리 고도의 이질성과 단위 불균형성을 내포하므로, 선형 모형의 설명력을 제약한다. 이에 따라, XGBoost와 같은 비선형 기계학습 기법을 적용함으로써 예측 정확도는 물론, 예측 변수의 영향력 해석이라는 두 가지 측면에서 실증적 유의성을 확보할 수 있다.

2. 선행연구

부동산 AVM은 가치평가의 객관성, 일관성, 그리고 효율성을 제고하기 위한 데이터 기반 계량적 추정기법으로 출발하여, 최근에는 머신러닝 알고리즘의 도입과 함께 지속적으로 발전해왔다.

배성완·유정석(2017)은 부동산 가격지수의 예측에 딥러닝 기법을 적용하고, 이를 전통적인 시계열 분석 방법과 비교함으로써 딥러닝의 활용 가능성을 검증하였다. 구체적으로 DNN(deep neural network), LSTM(long short-term memory)

모형을 활용하여 ARIMA(autoregressive integrated moving average) 모형과의 예측 성능을 비교하였다.

하대우 외(2019)는 코스피 200 주가지수의 등락 방향 예측을 위해 XGBoost 모형을 적용하고, 그 성능을 LSTM 및 자기회귀모형(AR)과 비교하였다. 분석 결과, XGBoost는 시계열 기반 주가 예측에서 구조적 유연성과 계산 효율성을 동시에 확보하며, 비선형성과 변동성이 높은 데이터에서도 높은 예측 정확도를 보였다.

김상환(2022)은 XGBoost 알고리즘을 활용하여 주식시장의 향후 등락 방향을 분류하는 예측모형을 구축하였다. 예측 정확도 평가지표로는 오분류율을 사용하였으며, 평균적으로 약 45% 수준의 오분류율을 기록하였다. 이는 완전한 효율적 시장 가설하에서 기대되는 임계값인 50%를 하회하는 결과로, XGBoost 기반 모형이 일정 수준의 정보 예측력을 지닌 것으로 해석할 수 있다.

이선구 · 유선종(2024)은 LSTM 알고리즘을 활용하여 부동산 조각투자 자산의 가격 예측 가능성을 실증적으로 제시하였으며, 기술지표와 거시경제 변수를 포함한 모델이 높은 예측력을 보였다고 보고하였다.

김수아 외(2024)는 생성형 AI를 활용하여 뉴스 기사 본문을 요약하고, 감성 분석을 통해 도출한 뉴스 감성 지수를 부동산 가격 예측모형에 반영하였다. RMSE(root mean squared error)가 전반적으로 감소하였으며, 이는 텍스트 기반의 사회 · 경제적 정서 지표가 부동산 시장의 가격 변동 예측에 실질적으로 기여할 수 있음을 실증적으로 입증한 연구로 평가된다.

배성완 · 유정석(2018)은 아파트 매매실거래가격지수를 예측하기 위해 머신러닝과 시계열모형을 비교 분석한 결과, 시장 급변 시 머신러닝 기반 모형이 시계열모형 대비 우수한 예측 성능을 보였음을 실증하였다.

김규석 외(2024)는 부동산의 공간적 상관성을 반영한 시계열 딥러닝 기반 예측모형을 제안하고, 서울시 3,000세대 이상 아파트 단지를 대상으로 RNN(recurrent neural network), LSTM, GRU(gated recurrent unit) 세 가지 모델의 성능을 비교하였다. 본 연구는 공간적 상호작용을 고려한 딥러닝 기반 접근법이 대규모 단지 아파트 가격 예측에 실질적인 효용을 가질 수 있음을 시사하였다.

신은경 외(2021)는 지역 간 가격 방향성이 유사한 군집을 SOM으로 분류한 후, 해당 클러스터의 지역 정보를 LSTM 모델에 입력하여 부동산 가격을 예측하였다. 실증 결과, LSTM이 SVR보다 예측력이 우수하였으며, 특히 장기 예측(3개월 후)에서 더 높은 성능을 보였다.

문혜정 · 조남욱(2024)은 2016~2023년 온비드에서 거래된 8,394건의 임야 공매 데이터를 바탕으로, 주성분 분석(principal component analysis, PCA)와 회귀분석을 통해 낙찰가격에 영향을 미치는 요인을 도출하고, SVR, XGB 등 다양한 머신러닝 기법을 적용하여 낙찰가를 예측하였다. 그 결과, SVR이 가장 우수한 성능(MAPE 4.17%)을 보였으며, 낙찰가격에 영향을 미치는 주요 변수로는 지가변동률, 입찰정보, 토지면적, 맹지 여부 등이 확인되었다.

Yazdani(2021)는 주택 감정의 인종 편향과

주관성을 줄이기 위해 머신러닝 및 딥러닝 기반 가격 예측 모형을 설계하고, 헤도닉 회귀모형과 비교하였다. 콜로라도 볼더시의 사례 분석 결과, 랜덤포레스트와 인공신경망이 더 높은 예측 정확도를 보이며 비선형적 변수 관계를 효과적으로 반영하였다. 이는 기존 감정 방식의 한계를 보완할 수 있는 자동화된 AVM 대안으로서 머신러닝의 활용 가능성을 시사한다.

Zhan et al.(2023)은 홍콩 부동산 거래 189만 건의 대규모 데이터를 기반으로, 베이지안 최적화를 활용한 하이브리드 주택가격 예측 모형(HBOS, HBOB, HBOT)을 제안하였다. 특히 HBOS-CatBoost 모델은 기존 XGBoost 및 ConvLSTM 기반 모델 대비 예측 오차(RMSE)를 각각 5.11%, 25.56% 낮추며 뛰어난 성능을 입증하였다.

Hasan et al.(2024)은 주택 특성, 위치, 이미지, 설명문 등 다양한 데이터를 통합한 멀티모달 딥러닝 기반 주택 가격 예측 모형을 제안하였다. 텍스트와 이미지 임베딩을 함께 활용해 학습한 결과, 단일 데이터보다 예측 정확도가 유의미하게 향상되었음을 실증하였다. 이는 실제 부동산 매물 정보를 반영한 복합 데이터 기반 예측 모델의 유효성을 제시한 연구이다.

Hernes et al.(2024)는 폴란드 브로츠와프 1차 주택시장을 대상으로, 주택 가격을 예측하기 위한 애플리케이션과 머신러닝 기반 모형(SLR, GBR, LASSO, RF 등)을 개발하였다. 스크래핑 기반 데이터 수집과 다양한 회귀 기법을 활용한 예측 정확도는 약 90% 수준으로 나타났으며, 실무 및 상업적 활용 가능성을 제시하였다(<표 1>).

기존 AVM 관련 연구들은 대부분 입력변수의 정규화를 전제로 하여 예측모형을 학습하고, 결과값을 RMSE, MAE(mean absolute error) 등의 표준 오차 지표로 평가하는 구조를 취해왔다.

이에 본 연구는 정규화 과정을 생략하고, 실거래가 단위의 원값을 그대로 사용한 상태에서 XGBoost 알고리즘을 학습시킴으로써, 결과값이 직관적 해석이 가능한 실거래 단위(원화)로 산출되도록 하였다. 이를 통해 AVM의 실무적 수용성을 높이는 동시에, 예측값과 실거래가 간의 직접 비교가 가능하도록 하였다.

또한, 본 연구는 기존 연구들이 간과해온 예측 적중률(hit rate)을 함께 산출함으로써, 감정평가나 담보 심사 등 실무 기준과의 정합성을 평가하였다. 이는 기존의 비율 중심 오류 지표를 보완하고, 모형의 실효성과 실전 적용 가능성을 보다 실질적으로 검증하는 평가 방식이라 할 수 있다.

마지막으로, 분석 대상 자산군으로 다세대주택과 같은 구조적 이질성이 높은 비정형 부동산을 선택함으로써, 기계학습 모형의 일반화 가능성과 적용 유연성을 검증하였다. 이는 아파트와 같이 표준화된 자산이 아닌 복잡한 시장을 대상으로 예측 모형을 적용하였다는 점에서, 기존 연구와 구별되는 실증적 기여라 할 수 있다.

Ⅲ. 연구모형

부동산 실거래가 예측에 있어 적용 모형의 선택은 대상 자산군의 구조적 속성과 데이터 특성에 대한 정합성을 전제로 이루어져야 한다. 머신러

〈표 1〉 연구자 및 내용

연구자	연구 내용	특징
배성완 · 유정석(2017)	딥러닝(DNN, LSTM)을 활용한 부동산 가격지수 예측 및 ARIMA와의 비교	딥러닝 VS 시계열 비교, 딥러닝 우수성 확인
배성완 · 유정석(2018)	머신러닝(SVR 등)과 시계열모형의 예측력 비교, 급변 시 시장 적응성 분석	시계열보다 머신러닝이 급변 상황에 강함
하대우 외(2019)	XGBoost 기반 주가지수 예측 및 모형 비교	예측력 · 효율성 우수, 시계열 예측모형의 실용성 제시
신은경 외(2021)	SOM으로 군집화한 지역 정보를 LSTM 입력으로 활용한 장기예측모형 제시	LSTM 장기예측에서 성능 우수, 군집 기반 입력 활용
김상환(2022)	XGBoost를 활용한 주요 주식 종목의 등락 방향 예측	오분류율 45% 수준, 효율적 시장 가설 하 예측 유의성 확보
이선구 · 유선종(2024)	LSTM 기반 부동산 조각투자 자산 가격 예측, 기술지표와 거시변수 포함	부동산 조각투자 자산 대상 LSTM 실증
김수아 외(2024)	뉴스 감성 지수를 생성형 AI로 도출 후 LSTM, VAR 예측모형에 반영	감성분석 기반 사회적 지표 예측력 향상 확인
김규석 외(2024)	RNN, LSTM, GRU를 활용한 대단지 아파트의 공간적 예측력 분석	공간 상호작용 반영, 대단지 RNN 성능 우수
문혜정 · 조남욱(2024)	PCA 및 머신러닝을 이용한 임야 공매 낙찰가 예측, SVR 성능 우수	낙찰가격 주요 변수 도출 및 알고리즘별 비교
Yazdani(2021)	머신러닝 기반 예측모형이 기존 헤도닉 회귀보다 비선형 관계 설명에 우수	랜덤포레스트 · ANN이 헤도닉보다 예측 정확도 우수
Zhan et al.(2023)	Bayesian 최적화 기반 하이브리드 모델(HBOS 등), CatBoost가 RMSE 우수	CatBoost 기반 하이브리드모델의 정밀 예측 성능 입증
Hasan et al.(2024)	텍스트 · 이미지 임베딩을 활용한 멀티모달 딥러닝 주택가격 예측	텍스트+이미지+지리정보 통합 모델로 예측 정확도 향상
Hernes et al.(2024)	브로츠와프 시장 대상 예측 애플리케이션 및 머신러닝(GBR 등) 모델 개발	상업적 활용성 높은 모델, 약 90% 정확도

주 : DNN, deep neural network; LSTM, long short-term memory; ARIMA, autoregressive integrated moving average; RNN, recurrent neural network; GRU, gated recurrent unit; PCA, principal component analysis; RMSE, root mean squared error.

닝 기법에는 의사결정나무 기반의 전통적 모형부터, 순환신경망 구조의 LSTM, 집합학습 기반의 랜덤포레스트 및 그레이디언트 부스팅, 그리고 이들의 고도화 형태인 XGBoost와 LightGBM이 포함된다. 이 중 XGBoost는 반복적 부스팅 구조를 통해 변수 간 복잡한 비선형 상호작용을 효과적으로 포착하고, 이상치 및 이질성에 대한 강건성을 갖춘 것이 특징이다. 특히 부동산 실거래 데이터는 구조적 비표준성과 입지적 · 제도적 다양성이 내재되어 있어, 모델의 해석 가능성과 정밀 예측력이 요구된다. 이러한 맥락에서 XGBoost는 실거래 기반 AVM에 있어 이론적 적합성과 실

무적 유효성을 겸비한 대안으로 평가된다.

Chen and Guestrin(2016)이 발표한 논문에 따르면, 2015년 Kaggle에서 우승한 29개 과제 중 17개가 XGBoost를 활용하였으며, 이 중 8개는 XGBoost만을 단독으로 사용하였고, 나머지는 신경망 등과의 앙상블 형태로 활용되었다. 또한, Sharma et al.(2024)은 미국 에임스시 주택 데이터를 활용해 여러 회귀모형을 비교한 결과, XGBoost가 가장 높은 R^2(0.920)과 가장 낮은 MSE(0.015)를 보여 주택 가격 예측에서 가장 우수한 성능을 나타냈다고 밝혔다. Ibok(2025)는 1988~2020년 퍼스 부동산 데이터를 기반으로 PySpark와 Tableau를 활용해 주택 가격을 예측하였으며, XGBoost가 가장 높은 성능(R^2=80.7%)을 보여주었다. 성능 향상을 위해 로그 변환과 시계열 기반 접근이 제안되었다. 국내에 있어서는 부동산 투자 결정 예측 분석에 있어 XGBoost를 사용하였고(주현태, 2023) 이에 따라 본 연구원도 다른 알고리즘보다 뛰어나다고 평가받는 XGBoost를 선택하였다.

1. 예측모형의 이론적 기반

XGBoost는 Friedman(2001)의 GMB(gradient boosting machine)을 기반으로 발전된 알고리즘으로, 경사 하강법 기반의 순차적 학습과 2차 미분 기반 손실함수 보정을 통해 예측 정밀도와 학습 효율성을 동시에 달성할 수 있는 구조를 갖는다. 특히 본 알고리즘은 정규화 항을 활용한 과적합 방지 기능, 분산형 병렬 학습 구조, 결측값 자동처리 기능 등을 포함하고 있어, 현실 거래 데이터를 다루는 데 있어 높은 신뢰성과 확장성을 제공한다.

이러한 구조적 특성은 변수 간 비선형성, 상호작용 효과, 이질적 시계열성을 내재한 부동산 데이터에 효과적으로 대응할 수 있으며, 특히 다세대주택과 같이 물리적 구조 및 입지 여건이 표준화되지 않은 자산군의 가치평가에 있어 우수한 예측력을 기대할 수 있다.

2. 변수체계 및 입력구조 설계

연구모형의 독립변수는 실거래가에 영향을 미치는 것으로 판단되는 요소를 물리적 특성, 입지 및 환경요인, 거시경제 및 시장지표 등으로 구분하여 총 25개 독립변수를 선정하였다. 주요 변수는 〈표 2〉와 같다.

3. 학습구조 및 평가전략

모형 학습에는 시계열 구간을 고려한 계약연월 기반 분할법(time-based split)을 적용하였다. 이는 부동산 시장의 비정기적 거래 패턴과 정책·금융환경의 시계열 변동성을 반영하기 위한 구조로, 시계열적 종속성이 내재된 실거래 데이터를 보다 현실적으로 반영하는 학습 전략이다.

모형의 하이퍼파라미터는 기본값을 기준으로 n_estimators, max_depth, learning_rate 등의 성능 민감도를 조정하여 최적화하였으며, 학습 데이터와 테스트 데이터를 독립적으로 활용하였다(〈표 3〉).

〈표 2〉 변수내용

유형	내용	단위
종속변수	거래금액	만 원
물리적 속성 변수	전용면적	㎡
	대지권면적	㎡
	층수	층
	건축년도	년
	건축면적	㎡
	연면적	㎡
	승강기수	유/무
	세대수	세대 수
	지하철 거리	Degree(°)
	성범죄자 거리	Degree(°)
입지 및 환경 변수	지리좌표(x, y)	Degree(°)
	총 거주 인구수	명
	총 거주 세대수	명
	아파트 인구수	명
	직장 인구수	명
	아파트 세대수	세대 수
	평균 소득	만 원
	일평균 유동인구	명
시장 및 경제 변수	연립 다세대 매매 실거래가격지수(서울)	지수
	연립다세대 매매 실거래 평균가격(서울)	만 원/㎡
	연립다세대 매매 실거래 중위가격(서울)	만 원/㎡
	주택매매시장 소비심리지수(서울)	지수
	계절조정 연립 다세대 매매 실거래 가격지수(서울)	만 원/㎡
	소비자물가지수(서울)	지수
	시장금리(월)	%

〈표 3〉 학습구조 및 평가전략

유형	내용
학습 기간	2023년 11월~2024년 7월
검증(테스트) 기간	2024년 8월~2024년 10월
데이터 분할 방식	계약연월 기준 시계열 분할 (time-based split)
활용 알고리즘	XGBoost 회귀모형
하이퍼파라미터 조정	n_estimators, max_depth, learning_rate 등의 성능 민감도 최적화
평가지표	예측 적중률(hit rate)을 병행 산출하여 실무 정합성 평가

Ⅳ. 연구결과

1. 변수 출처

본 연구에서 활용한 실증자료는 다양한 공공 데이터베이스 및 공간정보 API를 기반으로 구축되었으며, 거래의 시점, 위치, 구조, 시장 요인을 다차원적으로 반영하였다.

먼저, 거래금액, 전용면적, 대지권면적, 층수, 건축년도 등은 국토교통부 실거래가 공개시스템(https://rt.molit.go.kr)을 통해 2023년 11월부터 2024년 10월까지 서울특별시 은평구에서 체결된 연립 · 다세대주택 매매 실거래자료(원 raw data 2,446건)을 수집하였다. 데이터는 계약일 기준으로 정렬되었으며, 가격 왜곡 가능성이 있는 개인 간 직거래 사례와 계약 해제 건은 분석 대상에서 제외하였다. 결과적으로, 부동산 중개업소를 통한 정상거래 1,839건을 최종 분석 대상으로 확정하였다.

건축면적, 연면적, 승강기 수, 세대수 등의 건

 https://doi.org/10.30902/jrea.2025.11.2.21

축물 물리적 속성 정보는 해당 주소지의 건축물대장 정보 중 '건축 HUB_건축물대장 표제부 조회' 오픈 API(공공데이터포털)를 통해 확보하였다.

지하철 거리, 성범죄자 거리, 지리좌표(x, y)는 '도로명주소 기반 공간정보 변환 API'를 활용하여 각 거래 주소를 중심으로 직접 계산하였으며, UTM-K 기준 좌표계로 통일하였다.

총 거주 인구수, 총 거주 세대수, 아파트 인구수, 아파트 세대수, 직장 인구수, 평균 소득, 일평균 유동인구 등은 해당 거래지점을 중심으로 반경 100m 내의 공간영역을 지정하여 수집하였다. 해당 데이터는 GIS(geographic information system) 기반 big data에서 자료를 수집하였다.

연립 다세대 매매 실거래 가격지수(서울), 중위가격, 평균가격, 주택매매시장소비심리지수, 계절조정 연립 다세대 실거래가격지수, 소비자물가지수는 KOSIS 국가통계포털을 통해 월별 데이터를 수집하였으며, 거래 시점에 병합되도록 시계열 정렬을 수행하였다. 마지막으로, 시장금리는 해당 기간의 한국은행 경제통계시스템(ECOS)에서 제공하는 월별 국고채 3년물 수익률 자료를 사용하였다(〈표 4〉).

〈표 4〉 변수출처

변수	출처
거래금액	국토교통부(2025)
전용면적	
대지권면적	
층수	
건축년도	
건축면적	행정안전부(2025a)
연면적	
승강기수	
세대수	
지하철 거리	행정안전부(2025b)
성범죄자 거리	
지리좌표(x, y)	
총 거주 인구수	BIZ-GIS(2025)
총 거주 세대수	
아파트 인구수	
직장 인구수	
아파트 세대수	
평균 소득	
일평균 유동인구	
연립 다세대 매매 실거래 가격지수(서울)	통계청(2025)
연립다세대 매매 실거래 평균가격(서울)	
연립다세대 매매 실거래 중위가격(서울)	
주택매매시장소비심리지수	
계절조정 연립 다세대 매매 실거래 가격지수	
소비자물가지수	
시장금리	한국은행(2025)

2. 기초통계량

본 연구에서는 서울특별시 은평구에서 실제로 거래된 연립 · 다세대 주택 총 1,839건을 분석 대상으로 선정하고, 해당 부동산의 물리적 · 입지 · 시장 특성을 포괄하는 다양한 변수에 대해 기초통계량 분석을 실시하였다. 변수는 거래금액, 면적, 건축 정보, 경제지표, 공간좌표 및 주변 환경 데이터를 포함하며, 일부 변수는 외부 공간 통계 API를 활용하여 반경 100m 이내에서 정량화된 값을

수집하였다(〈표 5〉).

먼저, 거래금액의 평균은 약 28,428.89만 원이며, 최솟값 4,219만 원, 최댓값 99,000만 원으로 분포하고 있다. 전용면적은 평균 47.32㎡, 대

〈표 5〉 기술통계량

구분	최소값	최대값	평균	표준편차	분산
전용면적(㎡)	12	102	47.32	15.297	233.985
대지권면적(㎡)	3.33	783.00	32.14	41.29	1,704.73
층(충)	-1	11	3.00	1.929	3.721
건축 연도(연)	1968	2023	2006.69	10.537	111.022
지하철 거리(degree)	0.00581	0.07269	0.03450	0.01136	0.00013
연립 다세대 매매 실거래가격지수(서울)	133.1	141.9	137.604	2.9445	8.670
연립다세대 매매 실거래 중위가격(서울, 만 원/㎡)	662.3	771.1	707.088	35.7102	1,275.220
연립다세대 매매 실거래 평균가격(서울, 만 원/㎡)	774.6	898.8	819.837	41.4335	1,716.734
주택매매시장소비심리지수(서울)	99.6	122.1	111.217	6.6391	44.077
계절조정 연립 다세대 매매 실거래 가격지수(서울)	133.1	140.7	137.370	2.0416	4.168
성범죄자 거리(degree)	0.00011	0.30198	0.00668	0.01878	0.000
소비자물가지수(서울)	111.98	114.04	113.3232	0.57742	0.333
시장금리(월, %)	2.868	3.771	3.21639	0.241796	0.058
x_point(degree)	126.89334	127.20343	126.91825	0.01951	0.000
y_point(degree)	37.42326	37.64415	37.60715	0.01388	0.000
승강기수(개수)	0	3	0.42	0.521	0.271
세대수(명)	0	35	10.23	3.849	14.817
건축면적(㎡)	13.00	785.91	156.78	71.59	5,124.99
연면적(㎡)	81.10	3,330.67	625.93	326.70	106,731.52
총 거주 세대수(명)	94.0	31727.0	624.685	745.8126	556,236.418
총 거주 인구수(명)	229.0	67,185.0	1,318.923	1,580.4704	2,497,886.619
아파트 세대수(세대수)	0	7,045	39.99	176.534	31,164.410
아파트 인구수(명)	0	16,923	87.86	419.625	176,084.957
직장 인구수(명)	0.0	17,395.0	209.589	444.4837	197,565.771
평균 소득(만 원)	4,891.0	8,282.0	5,865.858	367.7280	135,223.868
일평균 유동인구(명)	93.0	192,155.0	3,839.859	4,945.7492	24,460,435.449
거래금액(만 원)	4,219	99,000	2,8428.89	11,050.919	122,122,806.266

 https://doi.org/10.30902/jrea.2025.11.2.21

지권면적은 평균 32.14㎡로 나타났으며, 표준편차가 각각 15.30㎡, 41.29㎡로 나타나 실거래 주택 간 면적 차이가 큼을 보여준다. 층수는 평균 3층이며, 일부 반지하(-1층) 포함 물건도 존재하였다.

지하철 거리는 평균 0.03450으로, 대부분의 거래대상이 대중교통 접근성이 양호한 위치에 있음을 보여준다. 성범죄자 거리의 평균은 0.00668이며 표준편차는 0.01878로, 일부 지역에서는 민감한 환경 요소가 부동산 가치에 영향을 미칠 가능성도 제기된다.

연립다세대 실거래 평균가격(서울)은 819.84만 원/㎡, 중위가격은 707.09만 원/㎡이며, 계절조정 실거래가격지수는 평균 137.37로 집계되었다. 주택매매시장 소비심리지수(서울)는 평균 111.22, 표준편차 6.64로 나타나 투자심리 변동성이 존재함을 보여준다. 시장금리는 평균 3.22%로, 비교적 안정적 금리 수준에서 거래가 이루어졌음을 알 수 있다.

좌표 데이터(x, y)는 각각 평균 126.918, 37.607로 나타났으며, 표준편차가 매우 작아 공간적으로 국지화된 범위에서 수집되었음을 확인할 수 있다.

주변 환경 변수의 경우, 본 연구는 각 거래대상 부동산을 기준으로 반경 100m 내의 생활·상권·주거 데이터를 추출하여 변수로 구성하였다. 이에 따라 총 거주 세대수는 평균 624.69세대, 총 거주 인구수는 평균 1,318.92명으로 나타났다. 이는 해당 지역이 중밀도 이상의 주거 밀집지역이라는 점을 시사한다. 또한, 직장 인구수는 평균 209.59명, 일평균 유동인구는 평균 3,839.86명으로 집계되어, 실거래지 인근에 일정 수준의 업무·상권 기능이 존재함을 확인할 수 있다. 이처럼 유동인구 및 고정 주거인구는 향후 거래금액에 영향을 미치는 생활권 수요 지표로 기능할 가능성이 높다.

한편, 평균 소득(거주지 기준)은 5,865천 원 수준으로 나타났으며, 소득분포의 표준편차가 367.73천 원에 달하는 점을 고려할 때, 분석 대상 지역은 비교적 중상위 계층이 분포하면서도 소득 다양성도 공존하는 특징을 보인다.

3. 실증분석

연구는 개별 부동산 거래 단위를 하나의 인스턴스로 간주하고, 각 건별 실거래금액을 예측 대상으로 설정하였다. 입력 변수는 전용면적, 층수, 건축년도 등 물리적 속성과, 지하철 거리, 유동인구, 평균소득 등 입지·환경 변수, 그리고 실거래가격지수, 소비자물가지수, 시장금리 등 시장·경제 변수를 포함하여 총 25개로 구성되었다.

데이터는 총 1,839건으로, 2023년 11월부터 2024년 10월까지 서울특별시 은평구에서 체결된 연립·다세대 주택의 실거래 사례를 기반으로 한다. 이 중 계약년월을 기준으로 시계열 분할하여 2023년 11월부터 2024년 7월까지 체결된 1,272건을 학습 데이터로, 2024년 8월부터 10월까지의 567건을 테스트 데이터로 각각 구분하였다. 이와 같이 시점 기반의 분할(time-wise split)은 향후 시계열 구조를 내재한 거래 예측 모형의 확장 가능성을 열어두는 동시에, 데이터 누수(leakage)를 방지하며 예측 타당성을 제고하기 위한 전략이다.

또한 정적 회귀모형(static regression model)의 형태를 채택하였다(〈표 6〉).

XGBoost 회귀모형의 예측 성능을 극대화하기 위해 주요 하이퍼파라미터인 n_estimators, max_depth, learning_rate를 중심으로 실험을 구성하였다. 이들 매개변수는 트리 기반 부스팅 알고리즘의 복잡도와 일반화 능력을 결정짓는 핵심 요소로, 예측 정밀도와 과적합 방지 간의 균형을 조절하는 데 중요한 역할을 한다.

n_estimators는 전체 부스팅 반복 횟수를 설정하는 변수로, 값이 증가하면 모델의 세밀한 학습이 가능해지나 과적합 위험도 높아진다. 반면, 값이 너무 작으면 학습이 불충분해 과소적합 문제가 발생할 수 있다. 이에 본 연구는 실험적 접근을 통해 100, 300, 500의 값을 비교 적용하였다.

max_depth는 개별 결정트리의 최대 깊이를 제한함으로써 트리의 복잡성과 변수 간 상호작용 포착 능력을 조절한다. 깊이가 깊을수록 복잡한 패턴을 학습할 수 있지만 과적합 가능성이 커지며, 반대로 얕은 트리는 표현력이 떨어질 수 있다.

〈표 6〉 예측 구조

변수	출처
예측 단위	1건의 연립 · 다세대 주택 실거래 데이터
입력 변수 구성	물리적 속성, 입지 및 환경 변수, 시장 및 경제 변수
예측 대상 변수	실거래금액(단위: 만 원)
데이터 총량	1,839건(2023.11~2024.10)
학습 데이터셋	1,272건(2023.11~2024.07 계약 체결 분)
테스트 데이터셋	567건(2024.08~2024.10 계약 체결 분)
예측 구조	시계열 누적 기반 예측 아님. 각 건을 단일 인스턴스로 처리하는 정적 회귀모형 적용

김상환(2022)은 XGBoost 기반 주가예측 실증연구에서 깊이 3, 4, 5, 6을 실험적으로 비교하였으며, 본 연구에서는 이 선례를 참고하여 3, 5, 7로 설정하고 분석을 수행하였다.

또한, learning_rate는 각 부스팅 단계에서 새 트리의 기여도를 조절하는 변수로, 수렴 속도와 모델 안정성에 영향을 미친다. 낮은 학습률은 안정적인 수렴과 과적합 방지에 유리하지만 학습 시간이 길어지고, 반대로 높은 학습률은 빠른 수렴이 가능하나 예측의 불안정성을 유발할 수 있다. 김은미 외(2020)는 주택매도 결정요인 분석 및 예측모형 구축 연구에서 환경 변수에 적절한 학습률로 0.1, 0.01, 1, 0.2를 설정하여 실험한 바 있으며, 본 연구는 안정적 학습을 위해 0.1, 0.01, 0.001의 학습률을 적용하여 비교 분석하였다(〈표 7〉).

XGBoost 회귀모형 실험에서는 주요 하이퍼파라미터를 다음과 같이 설정하였다. 먼저 n_estimators는 부스팅 반복 횟수를 의미하며, 본 연구에서는 총 300회의 반복 학습을 수행하여 모델의 예측 정밀도를 제고하였다. max_depth는 개별 결정트리의 최대 깊이를 의미하며, 본 모형에서는 깊이 5로 설정함으로써 설명 변수 간 복잡한 비선형 상호작용을 효과적으로 학습할 수 있도록 설계하였다. 아울러 learning_rate는 각 반

〈표 7〉 하이퍼파라미터 값 설정

하이퍼파라미터	값
n_estimators	100, 300, 500
max_depth	3, 5, 7
learning_rate	0.1, 0.01, 0.001

 https://doi.org/10.30902/jrea.2025.11.2.21

복 단계에서의 학습 반영 비율로 설정되며, 본 연구에서는 0.1의 값을 적용하여 수렴 속도와 일반화 능력 간의 균형을 도모하였다.

이러한 설정을 기반으로 한 실증 결과, 훈련 데이터셋에서는 train score 0.018524로 매우 낮은 예측 오차를 나타냈으며, 테스트 데이터셋에서도 test score 0.136511 및 결정계수 R^2 0.747을 기록하였다. 특히 R^2가 0.747이라는 수치는, 본 모델이 실제 거래금액의 약 74.7%에 해당하는 변동성을 설명할 수 있음을 의미하며, 이는 고차원적 구조와 비선형성이 강한 부동산 실거래 데이터에 대해서도 본 회귀모형이 통계적으로 유의한 설명력을 갖추고 있음을 시사한다.

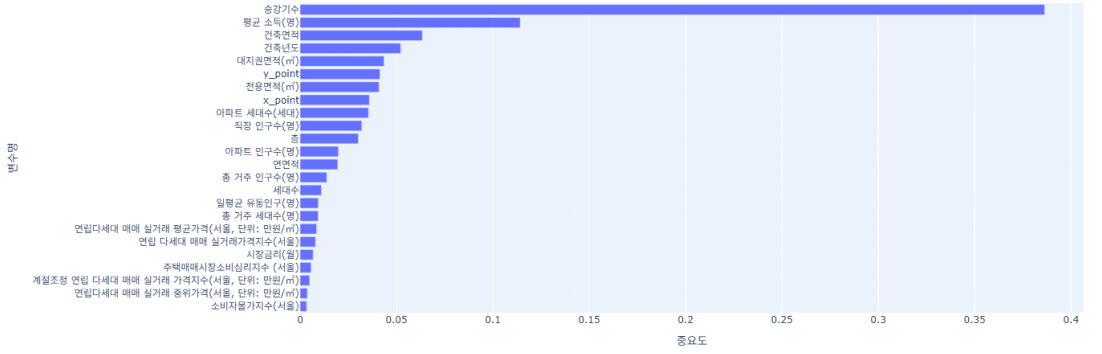
또한, MAPE는 13.6%로 산출되었으며, 이는 평균적으로 예측값이 실제 거래금액 대비 약 13.6% 수준의 상대 오차를 보였다는 것을 의미한다. 특히 본 연구의 테스트 데이터 구간에서 실거래가 평균은 약 29,163만 원(2억 9,163만 원)으로 나타났으며, 이를 기준으로 평균 예측 오차는 약 3,966만 원 수준으로 계산된다. 이는 AVM의 연립다세대 주택 실무 활용에 있어 충분히 수용 가능한 정밀도 수준으로 해석될 수 있으며, 가격대가 높은 고가 거래를 포함한 다양한 거래군에서도 일정 수준 이상의 일관된 예측 성능을 확보하였다는 점에서 실질적 의의가 있다(〈표 8〉).

〈그림 1〉은 XGBoost 회귀모형을 기반으로 산출된 예측 변수 중요도(feature importance)를 gain 기준으로 시각화한 결과이다. Gain은 각 변수의 분할이 모델의 예측 정확도 향상에 기여한 정도를 측정하는 지표로서, 본 분석에서는 거래금액 예측에 있어 각 독립변수가 갖는 기여도를 정량적으로 평가하였다.

분석 결과, '승강기 수'는 전체 변수 중 가장 높은 중요도를 보이며 약 38%의 기여도를 나타냈

〈표 8〉 예측 구조

R^2(결정계수)	n_estimators	max_depth	learning_rate	Train score	Test score
0.747	300	5	0.1	0.018524	0.136511



〈그림 1〉 변수중요도

다. 이는 다세대주택 유형에서 승강기의 존재 여부가 주거 선택의 핵심적 판단 요인으로 작용함을 시사한다. 특히 승강기는 고층 거주 시 이동 편의성을 결정짓는 핵심 인프라로, 설치 여부에 따라 동일 면적 대비 가격에 유의미한 차이를 발생시킨다. 최근 5층 이상 신축 비율이 증가하고 있는 도시형 주거지에서, 승강기의 존재는 물리적 편의성은 물론 고령층 및 가족 단위 거주자의 수요에도 부합하는 설비로 인식되고 있다.

다음으로 평균 소득은 약 11%의 중요도로 분석되었다. 이는 해당 주택의 입지 주변의 경제적 기반이 구매력과 선호도에 직접적인 영향을 미치고 있음을 보여준다. 소득 수준이 높은 지역일수록 생활 인프라와 교육 여건 등이 양호할 가능성이 높고, 이로 인해 동일한물리적 조건의 주택이라 하더라도 상대적으로 높은 가격 형성이 가능하다.

건축면적(6%)과 대지권면적(4%)은 물리적 자산의 규모와 구조를 나타내는 변수로, 거주 효율성과 장래 가치를 동시에 반영한다. 건축면적은 실내 공간의 활용 가능성과 직결되며, 대지권면적은 향후 재건축 또는 담보 가치 측면에서 중요한 판단 요소로 작용한다. 대지권이 넓을수록 토지 지분이 크다는 의미이며, 이는 장기적 자산가치 상승 기대에 긍정적 영향을 미친다.

건축 연도(5%)는 건물의 노후도와 직결되며, 신축 여부에 따라 건축 기준, 에너지 효율, 건물 관리 상태 등이 달라진다. 이러한 요소들은 실거래가 형성에 있어 실수요자들의 주관적 평가와 직결되며, 신축 주택에 대한 선호도가 높은 시장 특성을 반영한 결과로 해석된다.

전용면적(4%), 층(3%), 아파트 세대수(3%)는 주택 내부 구조와 쾌적성을 나타내는 지표이다. 전용면적은 실제 생활공간의 크기를, 층수는 조망권과 채광, 소음 민감도 등을 반영하며, 세대수는 단지의 규모나 관리 편의성과 관련이 있어 주택의 전반적인 인식에 영향을 미친다.

공간적 좌표인 x_point(3%)와 y_point(4%)는 지도상의 위치를 수치로 반영한 변수로, 명시적인 지역명 없이도 위치 정보를 모델이 인식하게 해주는 역할을 한다. 이러한 수치는 비선형 회귀모델에서 지역별 가격 패턴을 정량적으로 학습하는 데 유리하게 작용한다. 특히 정형화되지 않은 공간 변수의 비정형 특성을 효과적으로 반영한다는 점에서, 위치 데이터의 활용 가능성을 실증적으로 보여준다.

직장 인구수(3%)는 해당 지역의 배후 고용 수요를 나타내며, 생활권 내 자족 기능이나 출퇴근 편의성 등을 반영하는 변수이다. 직주근접성이 중요한 수도권 시장에서, 고용 밀집지역과의 근접성은 주택 수요를 유인하는 요인으로 작용하고 있다.

반면, 금리, 소비자물가지수, 주택시장 소비심리지수, 실거래가격 지수 등 거시경제 변수들은 상대적으로 낮은 중요도를 보였다. 이는 본 연구가 전국 단위의 시장 동향이 아닌, 개별 주택의 실거래가 예측에 초점을 맞추었기 때문으로 보인다. 이러한 변수들은 광역적 수준의 가격 흐름에는 유의미할 수 있으나, 개별 필지 단위의 미시적 거래 분석에서는 설명력이 제한될 수 있다. 이는 부동산 자동평가모형에서 거시지표의 기계적 삽입이 항상 높은 예측 성능으로 이어지지 않는다는 점을 시사하는 실증적 근거라 할 수 있다.

〈그림 2〉는 각각 테스트셋과 학습셋을 대상으로 한 XGBoost 회귀모형의 예측값과 실제 거래금액 간 산점도이다. 학습셋의 경우, 예측값이 기준선(y=x)을 따라 정렬되며 전 구간에서 높은 결정계수($R^2 \approx 1$)를 기록하는 등 높은 예측 정합성을 달성한 것으로 나타났다. 이는 모델이 학습 데이터의 구조를 효과적으로 파악하였음을 시사한다.

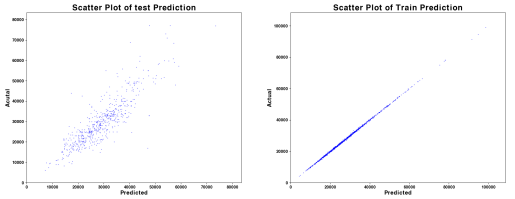
테스트셋의 산점도 역시 전체적으로 뚜렷한 선형성을 유지하고 있으며, 고가 거래 구간에서는 상대적으로 보수적인 예측값 분포가 관찰된다. 이러한 양상은 모델이 고가 영역에서 예측 안정성을 확보하며, 가격 이상치(outlier)에 대해 민감하지 않은 점진적 반응 특성을 가짐을 나타낸다. 이는 실무에서 AVM을 활용할 때 과도한 고평가를 방지하는 기능적 이점을 제공할 수 있다.

잔차 분석 결과도 이를 뒷받침한다. 학습셋 잔차는 중심을 기준으로 대칭성을 띠며, 오차 분산이 협소하게 수렴되어 전반적으로 예측 안정성이 확보된 양상을 보인다. 테스트셋 잔차 분포에서도 예측 오차는 전체적으로 무작위적이며, 고가 구간에서도 일정한 편차 범위 내에서 예측력이 유지되는 양상을 보인다. 이는 모델이 고가구간의 노이즈나 외생적 가격요인에도 견고한 예측 구조를 갖추고 있음을 시사한다(〈그림 3〉).

Ⅴ. 결론

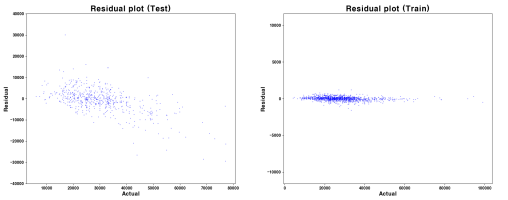
본 연구는 XGBoost 회귀모형을 기반으로, 서울특별시 은평구 다세대주택 실거래 데이터를 활용하여 부동산 AVM의 예측 성능을 실증적으로 분석하였다. 기존 AVM 연구에서 일반적으로 채택되었던 정규화 과정 없이, 원단위 실거래금액을 그대로 유지한 채 예측값을 도출함으로써, 실거래 단위에 근거한 직관적 · 실용적 예측모형의 구현 가능성을 제시하였다.

분석에 사용된 1,839건의 연립 · 다세대 실거래 중 1,272건은 학습용, 567건은 테스트용으로 구분되었으며, 물리적 속성, 입지환경, 거시경제변수 등 고차원의 설명변수를 통합하여 모델을 설계하였다. 실증 결과, 훈련 데이터셋에서는 매우 낮은 예측 오차(train score 0.018524)를 기록하였고, 테스트셋에서는 test score 0.136511, 결정계수 R^2 0.747, 평균절대백분율오차(MAPE) 13.6%를 보이며 일정 수준 이상의 일반화 성능을 확보하였다. 이는 전체 거래금액 변동의 약 74.7%



주 : 좌: test date, 우: train date.

〈그림 2〉 예측 vs 실제 데이터 산점도



주 : 좌: test date, 우: train date.

〈그림 3〉 잔차(residual) 분포도

를 본 모델이 설명하고 있으며, 약 86.4% 수준에서 실거래가를 근사하는 예측 정확도를 달성했음을 의미한다. 테스트 구간에서 실거래가 평균은 약 2억 9,163만 원으로 나타났으며, 이를 기준으로 평균 예측 오차는 약 3,966만 원(13.6%) 수준으로 계산된다. 특히 최근 보도된 부동산플래닛의 AI 기반 AVM은 중소형 상업용 부동산을 대상으로 약 80% 수준의 예측 정확도를 보였으며(한명현, 2024), 부동산R114 역시 AI 시세 시스템을 통해 기존 70%대에서 80% 수준으로 정확도를 개선한 바 있다(조용훈, 2025). 이러한 최근 기사 사례와 비교하였을 때, 본 연구의 예측 성과 86.4%의 근사 정확도는 학문적 기여뿐 아니라 실무적 활용 측면에서도 유의미한 결과로 판단된다.

더불어, 예측값과 실제값 간 산점도 분석 결과, 학습 데이터는 기준선을 중심으로 고도로 일치하는 분포를 보였고, 테스트 데이터 또한 전체적으로 명확한 선형관계를 유지하면서 고가 거래구간에서는 보수적 예측 특성이 확인되었다. 이는 모델이 고가 이상치에 과도하게 민감하지 않도록 조정 학습되었으며, 예측 안정성을 유지하면서도 실무적 해석 가능성을 제고한 결과로 평가된다.

잔차 분석에서도 학습셋은 중심을 기준으로 분산이 작고 대칭적인 구조를 유지하였으며, 테스트셋 역시 예측 오차가 무작위적이고 통계적으로 안정된 분포를 나타냈다. 특히 잔차의 편향이 크지 않으며, 고가 구간에서도 모델이 예외값에 대해 구조적 예측력을 유지하고 있음을 보여주었다. 이는 XGBoost가 다차원적 변수 상호작용 및 이질적 데이터 구조를 효과적으로 학습하고 반영하는 능력을 가졌음을 뒷받침한다.

이와 같은 결과는 XGBoost 기반 회귀모형이 부동산 실거래 데이터를 활용한 AVM 구현에 실증적으로 타당하다는 점을 확인시켜줄 뿐 아니라, 정규화 생략이라는 구조적 실험을 통해 기존의 수치화 중심 모델링이 갖는 해석의 한계를 보완하며, 예측 결과의 직관성과 실무 적용성을 동시에 확보할 수 있음을 입증한 것이다. 이는 감정평가 실무, 금융기관 대출심사, 정책적 모니터링 등 다양한 분야에서 보조 수단으로써 AVM 활용 가능성을 확장하는 데 기여할 수 있다.

향후 연구에서는 지역 간 공간이질성과 시간적 외생성 변동성을 반영할 수 있도록 공간회귀모형, 시계열 기반 딥러닝 모형(LSTM, transformer 등)과의 비교연구가 병행될 필요가 있다. 주요 하이퍼파라미터에 대한 자동화 최적화 기법의 도입을 통해 모델 성능의 정밀도를 향상시키고, 예측 모형의 설명력을 고도화하는 전략적 접근이 요구된다.

또한, 본 연구는 서울시 은평구를 공간적 범위로 설정하여, 다세대 및 연립주택을 중심으로 부동산 가치 변동의 구조적 특성을 실증적으로 분석하였다. 그러나 분석 대상지를 단일 자치구로 한정함에 따라, 지역 간 비교를 통한 일반화에는 일정한 제약이 존재한다는 점에서 한계가 있다. 특히 은평구는 연구기간 실거래가가 집계된 바 있으나, 이는 분석 가능한 최소 요건을 충족하는 수준으로서, 보다 정교한 모형 학습을 위해서는 추가적인 데이터 확보가 요구된다. 향후에서 유사한 거래량을 보이는 광진구, 송파구, 강북구, 중랑구 등 자치구를 추가 분석 대상으로 확대함으로써, 지역 간 구조적 특성 차이에 따른 가격 예측력 및

 https://doi.org/10.30902/jrea.2025.11.2.21

모형 적합성의 일반화 가능성을 검토하고자 한다. 이를 통해 다세대 · 연립주택 중심의 지역 기반 AVM 구축에 보다 실질적이고 보편화된 정책적 · 실무적 함의를 제공할 수 있을 것으로 기대된다.

결론적으로, 본 연구는 XGBoost 회귀모형을 활용한 정규화 생략 기반의 실거래 예측모형이 예측 정밀도, 해석 가능성, 실용성 측면 모두에서 현실 적합성이 높고, 시장 기반 AVM 구현에 있어 학술적 · 정책적 · 실무적 기여가 가능한 구조임을 실증적으로 제시하였다는 데 그 의의가 있다.

ORCID

이선구 https://orcid.org/0009-0005-5842-787X

참고문헌

1. 국토교통부. (2025). *국토교통부 실거래가. https://rt.molit.go.kr*
2. 김규석, 김경민, 조재우. (2024). 공간적 상관성을 고려한 딥러닝 기반 부동산 가격 예측 방법 제안. *한국정보기술학회논문지, 22*(1), 9-22.
3. 김상환. (2022). Xgboost 모형의 주가예측성과에 대한 실증연구. *사회과학연구, 39*(1), 29-55.
4. 김수아, 권미주, 김현희. (2024). 생성 AI기반 뉴스 감성 분석과 부동산 가격 예측: LSTM과 VAR 모델의 적용. *정보처리학회 논문지, 13*(5), 209-216.
5. 김은미, 김상봉, 조은서. (2020). 기계학습을 활용한 주택매도 결정요인 분석 및 예측모델 구축. *지적과 국토정보, 50*(1), 181-200.
6. 문혜정, 조남욱. (2024). 머신러닝 기반 한국 임야 공매의 낙찰가격 예측. *지능정보연구, 30*(2), 177-194.
7. 배성완, 유정석. (2017). 딥러닝을 이용한 부동산가격지수 예측. *부동산연구, 27*(3), 71-86.
8. 배성완, 유정석. (2018). 머신 러닝 방법과 시계열 분석 모형을 이용한 부동산 가격지수 예측. *주택연구, 26*(1), 107-133.
9. 신은경, 김은미, 홍태호. (2021). SOM과 LSTM을 활용한 지역기반의 부동산 가격 예측. *정보시스템연구, 30*(2), 147-163.
10. 이선구, 유선종. (2024). LSTM 모형을 활용한 부동산 조각투자 가격 예측. *부동산학연구, 30*(2), 25-43.
11. 조용훈. (2025.03.19.). *AI가 부동산 시세 예측한다… 부동산R114 "정확도 80% 달성".* 다음뉴스. *https://v.daum.net/v/20250319180237398*
12. 주현태. (2023). XGBoost를 이용한 부동산 투자 결정 예측 분석. *부동산분석, 9*(3), 55-69.
13. 통계청. (2025). *KOSIS 국가통계포털.* https://kosis.kr/index/index.do
14. 하대우, 김영민, 안재준. (2019). XGBoost 모형을 활용한 코스피 200 주가지수 등락 예측에 관한 연구. *한국데이터정보과학회지, 30*(3), 655-669.
15. 한국은행. (2025). *한국은행 경제통계시스템.* https://ecos.bok.or.kr
16. 한명현. (2024.11.05.). *AI 활용해 입지분석 가능한 자동평가모델 개발.* 집코노미. https://www.hankyung.com/article/2024110520161?utm_source=chatgpt.com
17. 행정안전부. (2025a). *건축HUB_건축물대장 정보서비스 API.* https://www.data.go.kr/index.do
18. 행정안전부. (2025b). *도로명주소 기반 공간정보 변환 API. https://www.juso.go.kr/openIndexPage.do*
19. BIZ-GIS. (2025). *GIS기반 Big Data. http://bigdata.biz-gis.com/*
20. Chen, T., & Guestrin, C. (2016). Xgboost: Ascalable tree boosting system. In *Proceedings of*

the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). Association for Computing Machinery.

21. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232.
22. Hasan, M. H., Jahan, M. A., Ali, M. E., Li, Y. F., & Sellis, T. (2024). *A multi-modal deep learning based approach for house price prediction.* https://arxiv.org/abs/2409.05335
23. Hernes, M., Tutak, P., & Siewiera, M. (2024). Prediction of residential real estate price on primary market using machine learning. *Procedia Computer Science, 246*, 3142–3147.
24. Ibok, B. (2025). *Predicting property prices in perth using big data analytics: A PySpark and Tableau approach.* Coventry University.
25. Sharma, H., Harsora, H., & Ogunleye, B. (2024). An optimal house price prediction algorithm: XGBoost. *Analytics, 3*(1), 30–45.
26. Yazdani, M. (2021). *Machine learning, deep learning, and hedonic methods for real estate price prediction.* https://arxiv.org/abs/2110.07151
27. Zhan, C., Liu, Y., Wu, Z., Zhao, M., & Chow, T. W. S. (2023). A hybrid machine learning framework for forecasting house price. *Expert Systems with Applications, 233*, 120981.

논문접수일: 2025년 5월 11일
심사(수정)일: 2025년 6월 30일
게재확정일: 2025년 7월 18일

국문초록

본 연구는 서울시 은평구 다세대주택 실거래 데이터를 기반으로, XGBoost 회귀모형을 활용하여 부동산 자동가치산정모형(automated valuation model, AVM)의 예측 성능을 실증적으로 검토하였다. 기존 Min-Max 정규화 과정을 하지 않고, 실거래 금액의 원단위를 유지한 상태에서 예측값과 실제값을 직접 비교함으로써, 해석의 직관성과 실무 활용성을 동시에 확보하고자 하였다. 2023년 11월부터 2024년 10월까지 1년간의 연립·다세대 실거래 1,839건 중 1,272건은 학습용, 567건은 테스트용으로 사용되었으며, 물리적 속성, 입지, 환경 요소, 시장 및 거시경제 지표를 포함한 변수를 기반으로 예측모형을 구성하였다. 분석 결과, 테스트 데이터 기준 test score 0.136511, R^2는 0.747, mean absolute percentage error는 13.6%로 도출되었으며, 이는 전체 거래금액의 약 86.4% 수준에서 실거래가를 근사 예측한 결과로 해석된다. R^2 0.747은 거래금액의 약 74.7%를 본 모델이 설명하고 있음을 의미하며, 테스트 구간 실거래가 평균은 약 2억 9,163만 원으로 나타났으며, 이를 기준으로 평균 예측 오차는 약 3,966만 원 수준으로 확인됐다. 이는 머신러닝 기법의 적용 가능성과 향후 정교한 부동산 가치평가모형 구축에 있어 학문적 및 실무적 의의를 동시에 가진다.

주제어 : 부동산 자동가치산정모형(automated valuation model, AVM), XGBoost, 실거래가 예측, 머신러닝, 다세대주택