



머신러닝 기반 주택공급량 추정모형의 개발과 성능 비교 - 단순 평균법과 생존분석 방법과의 비교를 중심으로 -

Developing a Machine Learning-Based Housing Supply Estimation Model

- A Comparative Analysis with Simple Averaging and Survival Analysis -

박태이* · 주승민**

Taeyi Park · Seung-Min Joo

■ Abstract ■

This study develops and evaluates models for predicting the realized housing supply from apartment projects permitted in the Seoul Metropolitan Area. Using permit histories from January 2010 to December 2019 as training data and apartment projects permitted in 2020 as a holdout test set, the study compares a simple averaging approach, a survival analysis model, and a Light Gradient Boosting Model (LightGBM) 4-way ensemble model. Whereas the benchmark models rely on static information available at the time of permit approval (forecasting), the proposed machine learning model is designed as a conditional short-term prediction (nowcasting) framework that incorporates dynamically updated administrative information. To isolate the predictive value of administrative supply data, the explanatory variables are limited to supply pipeline- and schedule-related information, including stage-specific backlogs and elapsed-time variables, while excluding market, financial, and spatial variables. The results show that the machine learning-based model, benefiting from this expanded information set, produces lower prediction errors in monthly completion forecasts and more accurately tracks cumulative completions than the benchmark models. These findings suggest that realized housing supply is explained less by aggregate permit volume than by the dynamically evolving backlog and duration structure embedded in the housing supply pipeline. Overall, the study demonstrates that a practically useful conditional prediction framework can be developed using administrative housing supply data alone.

Keywords: Realized housing supply, Completion volume forecasting, Survival analysis, Machine learning, Light gradient boosting model (LightGBM) ensemble

* 경북대학교 지리학과 석사(주저자) | Master, Department of Geography, Kyungpook National University | First Author | hahaha8875@naver.com |

** 한국부동산원 부동산분석처 거래분석부 부장(교신저자) | Director, Transaction Analysis Division, Real Estate Market Analysis Department, Korea Real Estate Board | Corresponding Author | joo@reb.or.kr |

1. 서론

주택공급은 주택가격 안정, 주거복지, 도시기반 시설 계획, 건설경기, 지역경제에 밀접하게 연결되는 핵심 정책변수이다. 이준용·손재영(2013)은 지역 단위의 물리적 주택공급 속도가 주택시장 변동을 설명하는 유의미한 요인임을 실증적으로 제시하였다. 그러나 정책 현장에서 일반적으로 활용되는 공급지표는 인허가, 착공, 분양, 준공 등 여러 단계의 통계로 분절되어 있어, 특정 시점에 실제로 어느 정도의 공급이 시장에 도달할 것인지를 직접 보여주지는 못한다. 동일한 인허가 물량이라 하더라도 사업장별 착공 시점, 공정 속도, 준공 지연 여부에 따라 실제 공급실현 시점은 크게 달라질 수 있기 때문이다.

이러한 문제의식은 국내외 선행연구에서도 반복적으로 제기되어 왔다. Lee et al.(2025)는 주택공급 과정의 주요 단계 사이에 시간적 불일치가 존재하며, 이러한 시차를 고려하는 것이 공급정책의 효과를 정확히 평가하는 데 중요하다고 지적하였다. 따라서 주택공급량 연구에서는 인허가 이후 실제 공급으로 이어지는 경로를 직접 추적·예측하는 접근이 필요하다.

해외 관련 문헌에서도 주택 공급물량은 정태적 물량이 아니라 동태적 파이프라인¹⁾으로 이해해야 한다고 보았다. Mayer and Somerville (2000)은 신규 주택건설 변화율에 반응하는 동태적 공급모형을 제시하였다. Green et al.(2005)는 도시별 공급탄력성이 크게 상이함을 보였다고

분석하였고, Saiz(2010)는 지리적 제약과 규제적 제약 등이 공급탄력성의 핵심 결정요인이라고 설명하였다. 이들 연구 결과는 공급이 단순 추세나 단순 평균 소요 기간 수준으로 설명되기 어려운, 비선형적이고 시간 지연적인 과정임을 시사한다.

따라서 주택공급량 추정 연구에서는 인허가 이후 실제 공급으로 이어지는 경로를 직접 예측하는 접근이 필요하다. 특히 정책적으로 중요한 질문은 “특정 기간에 허가된 물량이 이후 어느 기간에 얼마만큼 실제 준공되는가”이지, 단순히 “특정 기간에 몇 호가 허가되었는가”가 아니다.

최근 인공지능(AI)과 기계학습 기법의 발달로 학계에서 이를 접목한 다양한 실증 연구가 활발히 진행되고 있다. 그러나 그간의 인공지능 연구 동향을 살펴보면 주택 공급과 관련된 논의보다는 상대적으로 가격 예측, 거래량 분석, 자동가치산정(automated valuation model, AVM), 지수 산정 등에 편중되어 온 것이 사실이다. 구체적으로 홍정의(2021)는 랜덤 포레스트(Random Forest) 기반의 주택 대량평가모형을 제시하였고, 박대현 외(2021)는 기계학습 기반 주택시장 조기경보체계를 구축하였다. 또한 김이환 외(2022)는 기계학습을 활용한 아파트 매매가격지수 산정 방식을 제안하였으며, 이선구(2025)는 XGBoost(eXtreme gradient boosting) 기반의 부동산 자동가치산정모형을 실증적으로 분석한 바 있다.

이러한 선행연구들은 부동산 분야에서 Machine Learning 기술의 폭넓은 활용 가능성을 입증했다는 점에서 의의가 크다. 하지만 동시에 첨단 분

1) 파이프라인 기반 예측은 단일 시점의 정태적 예측이 아니라, 기준시점까지 누적된 행정 진행정보를 반영한 동태적 예측을 의미한다.

석 기법을 활용한 연구의 중심이 여전히 ‘가격 및 가치 평가’ 영역에 머물러 있으며, 시장의 근원적 지표인 ‘공급 실현량 추정’에는 적극적으로 활용되지 못하고 있다. 이에 본 연구는 기존 연구들에서 상대적으로 논의가 부족했던 주택 공급량 예측 분야에 초점을 두고, Machine Learning 기반 예측기법의 적용 가능성과 성능을 비교·검토하고자 한다.

본 연구는 2010년 1월 1일부터 2019년 12월 31일까지 인허가된 수도권 아파트를 학습자료로 설정하고, 2020년 1월 1일부터 2020년 12월 31일까지 인허가된 수도권 아파트를 테스트자료로 설정한다. 이는 2020년 이후 인허가 물량의 경우 현재 시점에서도 상당수가 아직 준공에 도달하지 않았을 가능성이 높기 때문에, 외표본 검증에 활용할 수 있는 자료 중 2020년 인허가 데이터가 가장 현실적인 평가대상이라고 판단하였기 때문이다. 이들 자료를 활용하여 단순평균법(simple averaging method), 생존분석 모형(survival analysis), Machine Learning 기반 예측모형을 이용해 공급 실현량 추정 방법을 설계한 뒤 2020년 1년간 인허가 건을 활용하여 준공 추정물량을 산정하고자 한다. 최종적으로 이를 2020년 인허가 물량의 실제 준공 실적과 비교하여 예측성능을 평가한다.

본 연구의 차별성은 다음과 같다. 첫째, 예측 방식과 대상의 전환이다. 예측 대상을 일반적인 공급 총량(total supply volume)이 아니라 특정 연도 인허가 집단의 향후 실현 경로(realization path)로 설정하였으며, 모형의 성격을 인허가 시점의 정태적 사전 예측(forecasting)이 아닌, 사업 진행에 따라 누적되는 행정 정보를 조건부로

활용하는 동적 단기 전망(nowcasting)으로 명확히 규정하였다. 둘째, 단순평균법과 생존분석을 기준모형으로 설정하고 Machine Learning 기반 모형의 추가 예측력을 외표본 자료를 통해 검증하였다는 점이다. 셋째, 시장·금융·공간 변수를 의도적으로 배제하고 공급 파이프라인 및 일정 정보만으로 예측체계를 구성함으로써, 공급행정 자료 자체의 설명력과 예측 가능성을 점검하였다는 점이다. 넷째, 일정기간의 준공물량과 누적 준공물량을 함께 비교하여 타이밍 정확성과 총량 정확성을 동시에 평가한다는 점이다.

본 논문의 구성은 다음과 같다. 제Ⅱ장에서는 주택공급의 실현 경로, 평균 기반 접근, 생존분석 기반 접근, Machine Learning 기반 부동산 예측연구를 검토하고 연구방법 설정한다. 제Ⅲ장에서는 분석데이터 전처리, 변수, 비교모형, 평가지표와 검증절차를 설명한다. 제Ⅳ장에서는 분석결과로서 2020년 인허가 건의 실제 준공 실적을 기준으로 모형별 예측성능을 비교한다. 제Ⅴ장에서는 연구결과를 요약하고 정책적 시사점, 연구의 한계 및 향후 과제를 제시한다.

II. 선행연구 검토

1. 주택공급의 실현 경로와 시간지연

주택공급은 허가와 동시에 실현되지 않는다. 인허가 이후 착공, 공정 진행, 준공에 이르기까지 상당한 시간이 소요되며, 일부 사업은 장기 지연되거나 취소되기도 한다. 따라서 공급정책의 성

과를 판단할 때에는 단순한 인허가 실적 자체보다 인허가 이후 실제 준공으로 이어지는 경로와 속도를 함께 고려해야 한다.

주택공급의 가격탄력성을 분석한 김형석·이영성(2008)의 연구에 따르면, 우리나라 주택시장의 공급 탄력성은 전반적으로 낮은 수치를 기록하며 비탄력적인 특성을 지닌 것으로 나타났다. 세부적으로 살펴보면, 비수도권 지역보다 수도권 지역의 공급이 수요 변화에 둔감하게 반응하였으며, 주택 유형 기준으로는 일반 주택 전체에 비해 아파트의 공급이 상대적으로 더 비탄력적인 양상을 띠는 것으로 분석되었다.

이건원 외(2018)는 2003년부터 2015년 사이 서울시 내 주택재개발사업지구를 대상으로 생존 분석 기법을 적용하여 사업 지연 요인을 규명하였다. 해당 연구는 정비구역 지정부터 관리처분인가에 이르는 전체 기간을 종속변수로 설정하였으며, 분석 결과 사업 대상지의 면적이 협소할수록, 그리고 계획된 임대 가구 수와 조합원의 규모가 클수록 사업 기간이 유의미하게 늘어나는 경향을 확인하였다.

정비사업뿐만 아니라 일반적인 신규 주택공급 과정에서도 소요 기간의 편차가 발생한다. 지규현 외(2017)는 수도권 주택사업을 중심으로 인허가 획득 후 실제 착공에 들어가기까지의 기간에 초점을 맞추어 생존분석을 수행하였다. 주택 유형별 비교에서는 아파트가 타 주택 유형보다 인허가 이후 착공까지 소요되는 기간이 구조적으로 더 긴 것으로 확인되었다. Lee et al.(2025)은 우리나라 주택건설사업 자료를 이용하여 절차 간 시간차를 분석하였고, 특히 현재 관측되는 준공이나

입주 지표가 상당 부분 과거 정책과 과거 시장조건의 결과라는 점을 강조하였다. 이는 현재의 허가실적과 미래의 실현공급량 사이에 구조적인 시차가 존재함을 보여준다.

해외 문헌에서도 공급의 시간구조는 핵심적인 분석대상으로 다루어진다. Mayer and Somerville(2000)은 주택건설을 가격 및 비용 변화에 대한 동태적 반응으로 해석하였고, 공급이 단기간에 완전히 조정되지 않는다는 점을 실증적으로 제시하였다. Green et al.(2005)은 도시권별 공급탄력성이 이질적임을 보였으며, Saiz(2010)는 지역적·규제적 제약이 공급반응의 차이를 만들어내고 설명하였다. 이러한 연구는 주택공급을 단순 평균이나 일률적 시차로 설명하기보다, 시간지연과 지역 이질성을 동시에 고려해야 함을 보여준다.

Zhang and Miller(2023)는 주택공급 과정을 주택 착공(housing starts)과 준공(completions)의 두 단계로 구분하고, 준공까지 소요되는 기간 분포를 콕스 비례위험모형(Cox proportional hazard model)으로 분석하였다. 이들은 공급모형에서 공급의 착공 시점뿐 아니라 착공 이후 실제 준공까지의 소요기간도 그에 못지않게 중요(equally important)하다고 보았으며, 주택 준공을 개별 프로젝트 단위의 생존시간(survival time)으로 다루었다. 해당 연구는 본 연구가 인허가 이후 준공까지의 소요기간을 예측대상으로 설정하는 데 중요한 선행연구로 볼 수 있다.

2. 평균 소요기간 활용 예측

예측문제에서 가장 기본적인 접근은 과거 관측

치의 평균이나 유사 사업의 평균 준공기간을 활용하는 것이다. 이러한 방식은 자료 요구량이 적고 해석이 직관적이라는 장점이 있으나, 구조적 변화나 비선형 관계를 충분히 반영하기 어렵다는 한계를 가진다. 그럼에도 단순평균법은 예측연구에서 중요한 기준모형으로 기능한다.

Makridakis and Hibon(2000)의 M3-competition는 복잡한 예측기법이 반드시 단순한 예측방법을 능가하는 것은 아님을 보여주었으며, 예측모형의 성능은 기준예측치와의 비교를 통해 평가될 필요가 있음을 시사하였다.

이러한 관점은 부동산 공급예측에도 적용될 수 있다. Machine Learning 모형이나 생존분석 모형이 더 정교한 구조를 갖는다고 하더라도, 그것이 곧바로 우수한 예측성능을 보장하는 것은 아니다. 따라서 해당 모형들이 실제 외표본(out-of-sample) 예측에서 단순평균법을 유의미하게 상회하는 성과를 보일 때에만 실질적인 방법론적 기여를 주장할 수 있다. 이에 본 연구는 단순평균법을 형식적인 비교대상이 아니라, 가장 기초적이고 재현 가능한 기준모형으로 설정한다.

3. 생존분석 기반 접근

본 연구의 맥락에서 생존분석(Survival Analysis)은 인허가 이후 준공이라는 사건(event)이 발생하기까지의 소요기간을 분석하는 데 활용될 수 있다. 특히 관측기간 종료 시점까지 준공되지 않은 사업을 우측 검열(right-censored) 자료로 처리함으로써, 미준공 사업을 단순히 제외할 때 발생할 수 있는 정보 손실과 추정 편향을 완화할 수 있

다(Kleinbaum and Klein, 2012). 주택공급 메커니즘에 이를 대입해 보면, 인허가 혹은 착공 이후 주택의 '준공'은 전형적인 사건이며, 그때까지 소요되는 기간은 '생존 시간(survival time)'으로 해석될 수 있다. 이와 관련하여 Cox(1972)가 제안한 비례위험모형(Proportional Hazards Model)은 기저위험함수(baseline hazard function)를 비모수적으로 유지하면서도 다양한 독립변수의 효과를 추정할 수 있어, 공급 지연 요인이나 준공 확률을 분석하는 데 매우 적합한 방법론으로 평가받는다.

국내외 선행연구들은 주택공급의 소요기간을 생존분석의 관점에서 분석함으로써, 해당 방법론이 공급기간 예측에 유용하게 활용될 수 있음을 보여준다.

국내의 경우, 지규현 외(2017, 2019)는 생존분석을 활용하여 수도권 주택사업의 인허가부터 착공, 준공에 이르는 단계별 소요 기간에 미치는 영향 요인을 분석하였다. 해당 연구들은 주택공급 기간이 고정된 상수가 아닌 시장 환경에 따라 변화하는 확률적 변수임을 강조하였다. 해외 문헌에서도 유사한 접근이 확인된다. Zhang and Miller(2023)는 개별 개발 프로젝트 수준에서 준공까지 소요되는 기간을 Cox 모형으로 분석하여 준공 확률(completion rate)에 미치는 영향을 규명하였다. 이들 선행연구가 공통으로 시사하는 바는, 정교한 주택공급 예측을 위해서는 단순한 총량적 추산(rule of thumb)을 넘어 개별 프로젝트의 준공 확률 분포를 추정하는 미시적 접근이 필요하다는 점이다.

본 연구는 이러한 선행연구의 논리를 수용하

여, 2020년 인허가 물량 집단(cohort)의 향후 준공 경로를 예측하기 위한 비교 모형으로 생존분석을 채택하였다. 2020년 인허가 건 중 분석 시점 현재까지 준공되지 않은 사례를 우측 검열 자료로 자연스럽게 처리할 수 있다는 점에서, 생존분석은 인허가-준공 간의 시차(time-lag)를 모형화하는 데 있어 이론적으로 가장 타당한 기준점이 될 수 있다.

그러나 주택 공급량의 총량적 예측 측면에서 생존분석은 몇 가지 방법론적 한계를 내포하고 있다. 첫째, 개별 단위 프로젝트의 위험률(hazard)을 정교하게 추정하더라도, 이를 특정 시점(월별 등)의 총량적 공급 실적(total volume)으로 집계하는 과정에서 나타나는 계절성, 특정 시기의 공급 집중 현상, 변수 간의 비선형적 상호작용 등을 모형 내에 충분히 반영하기 어렵다. 둘째, 주택공급 파이프라인 상의 단계별 잔여 물량과 경과 기간이 결합하여 만들어내는 복잡하고 동적인 수급 패턴은 선형적이거나 반모수적인(semi-parametric) 구조 가정만으로는 그 변동성을 포착하는 데 한계가 존재한다.

4. Machine Learning 기반 예측

국내 부동산 분야에서 인공지능과 기계학습 기법은 주로 주택의 가격 예측 및 가치 평가 영역을 중심으로 확산되어 왔다. 대표적으로 홍정의(2021)는 Random Forest 알고리즘을 활용해 주택 대량평가모형을 제시하였으며, 박대현 외(2021)는 기계학습 기반의 주택시장 조기경보체계를 구축하였다. 김이환 외(2022)는 기계학습을 접목한

새로운 아파트 매매가격지수 산정 방식을 제안하였고, 이선구(2025)는 XGBoost 기반의 AVM을 실증적으로 분석하였다. 이러한 선행연구를 종합하면 기계학습이 부동산 시장 데이터가 지닌 비선형성과 이질성을 효과적으로 포착할 수 있음을 입증하였다.

해외 연구의 흐름 역시 이와 유사한 궤를 같이한다. Pita et al.(2025)의 연구에 따르면, 부동산 분야의 기계학습 연구는 대체로 주택 가격 산정(house pricing) 문제에 집중되어 있으며, 방법론 측면에서는 Random Forest, Gradient Boosting, XGBoost 등이 주력으로 활용되고 있음을 지적하였다. 나아가 해당 연구는 기계학습 적용 시 발생할 수 있는 데이터 편향, 모형의 설명 가능성 확보, 그리고 객관적인 비교 기준(benchmark) 설정의 중요성을 강조하였다. 이는 인공지능과 기계학습기법을 주택 공급량 예측에 도입할 경우에도, 단순히 알고리즘의 우수성에 기대기보다 적합한 데이터 구축과 검증 체계 마련이 필수적임을 시사한다.

최근 부동산 가격 및 시장국면 예측 연구에서는 Random Forest, XGBoost, LightGBM 등 다양한 머신러닝 기반 모형이 활용되고 있다. 이 중 LightGBM은 Gradient Boosting Decision Tree 계열의 고효율 알고리즘으로, GOSS (gradient-based one-side sampling)과 EFB (exclusive feature bundling)을 통해 대용량·고차원 자료에서 학습 속도와 메모리 효율성을 개선한 것이 특징이다. Ke et al.(2017)은 LightGBM이 기존 GBDT(gradient boosting decision tree) 구현 대비 학습 속도를 크게 개선하면서도

유사한 수준의 정확도를 유지할 수 있음을 제시하였다. 이러한 특성은 지역, 주택유형, 면적, 층, 거래시점 등 다수의 범주형·연속형 변수가 혼재된 부동산 자료 분석에 적합할 것으로 판단된다. Tapia et al.(2025)은 칠레 산티아고 주택시장을 대상으로 LightGBM과 공간조정 헤도닉 모형을 비교하고, SHAP(shapley additive explanations) 분석을 통해 면적, 지역 특성, 학교 수준, 욕실 수 등 주요 요인의 비선형 영향을 확인하였다. Lahmiri et al.(2023)은 주택가격 예측에서 부스팅 앙상블 회귀나무가 인공신경망 및 다변량 회귀 모형보다 우수한 성능을 보였다고 보고하였다. 또한 Chen et al.(2023)은 XGBoost와 LightGBM이 주택가격 예측에서 높은 정확도와 학습속도를 가진다고 정리하였다. 최근 머신러닝 기법 중 CatBoost는 범주형 변수 처리에 우수한 성능을 보이는 것으로 알려져 있으나, 본 연구에서 다루는 수도권 인허가 데이터의 대규모 구조화 속성과 결측치 처리의 유연성, 그리고 반복적인 하이퍼파라미터 튜닝을 위한 메모리 효율성 측면을 종합적으로 고려하여 LightGBM을 최종 알고리즘으로 채택하였다. 다만, CatBoost 등 최신 부스팅 기법과의 성능 비교는 향후 연구 과제로 남겨둔다.

5. 선행연구와 차별성

기존의 논의를 종합할 때, 본 연구는 다음과 같은 세 가지 측면에서 선행연구들과 뚜렷한 차별성

을 갖는다.

첫째, 예측 대상의 전환이다. 기존 인공지능과 기계학습 연구들이 주로 가격이나 가치 평가에 집중한 반면, 본 연구는 특정 연도에 인허가를 받은 집단이 이후 실제 준공에 이르는 공급 실행 경로를 추적 및 예측한다는 점에서 차별성이 있다.

둘째, 독립변수의 의도적 통제다. 시장, 금융, 공간 등 거시적 환경 변수까지 광범위하게 포함했던 기존 연구와 달리, 본 연구는 주택 공급 파이프라인(인허가-착공-준공) 및 일정 관련 정보만을 설명변수로 제한하였다. 이는 외부 변수 개입을 최소화하고 공급 행정 통계 자료 자체가 지닌 내재적 예측력을 순수하게 검증하기 위함이다.

셋째, 엄밀한 성능 검증 체계 구축이다. 기계학습 모형의 성능을 객관적으로 평가하기 위해, 전통적인 추정 방식인 단순평균법과 생존분석 모형을 비교 기준(benchmark)으로 설정하였으며, 실제 데이터의 실적과 비교를 통해 객관적으로 성능을 검증하고자 한다.

III. 연구설계

1. 분석대상과 자료구성

본 연구는 수도권²⁾ 지역의 세움터 자료를 활용하여 2010년 1월부터 2019년 12월까지 인허가, 착공, 준공된 아파트를 학습표본으로, 2020년 1

2) 수도권 아파트는 공급 파이프라인의 단계별 식별이 비교적 명확하고, 공급 지연 및 준공 실행 여부가 주택시장 안정과 밀접하게 관련된다는 점에서 본 연구의 분석대상으로 설정하였다. 비수도권 및 비아파트 유형을 포함할 경우 지역별 시장여건과 주택유형별 사업구조의 이질성이 커질 수 있으므로, 본 연구에서는 분석대상을 수도권 아파트로 한정하여 모형의 기초적 예측 가능성을 검토하였다.

월부터 2020년 12월까지 인허가된 아파트를 독립적인 검증표본으로 설정하였다. 이와 같은 연도 기준의 표본 분할은 과거에 관측 가능한 정보만을 이용하여 이후 허가건의 준공 경로를 예측하는 실제 정책환경을 재현하기 위한 것이다. 예측 모형의 평가는 학습에 사용되지 않은 후행 시점의 자료를 통해 수행하는 것이 바람직하며, 시간의 순서를 보존한 외표본 검증은 미래 정보의 유입에 따른 평가 왜곡을 줄인다는 장점을 가진다(Hansen and Timmermann, 2012; Hyndman and Athanasopoulos, 2021).

예측 정확도 평가는 2020년 인허가 표본을 대상으로 수행하였다. 구체적으로는 각 모형이 2020년 인허가 물량 데이터를 활용하여 추정한 월별 준공 시점 또는 준공확률을 준공예상 물량으로 집계한 후, 이를 실제 준공 물량과 비교하였다. 이는 허가 시점에서 확보 가능한 정보만을 활용하여 향후 공급의 실현 경로를 얼마나 정확하게 예측할 수 있는지를 검토하려는 본 연구의 정책적 목적과 직접적으로 부합한다. 한편, 전통적인 생존분석에서는 관찰 종료 시점까지 사건이 발생하지 않은 표본을 우측 검열(right-censored)로 처리하는 것이 일반적이다. 그러나 일부 표본이 구조적으로 사건을 경험하지 않을 가능성이 존재하는 경우에는, 모든 표본이 궁극적으로 사건을 경험한다는 암묵적 가정을 두는 표준 생존모형만으로는 충분하지 않을 수 있으며, 이 경우 cure model과 같은 대안적 접근이 논의된다(Cai et al., 2012; Hosmer et al., 2008).

본 연구는 개별 사업의 순수한 지속기간 자체를 추정하는 데 목적이 있는 것이 아니라, 분석시

점 이후 실제로 실현 가능한 주택공급량을 보다 현실적으로 예측하는 데 목적이 있다. 이에 따라 2015년부터 2019년까지 인허가된 물량 가운데 2026년 현재까지도 준공되지 않은 사례는 지연 상태(미준공)으로 인식하였다. 이는 장기간 미준공 상태가 지속된 사업을 모두 잠재적 미래 공급으로 간주할 경우, 실현 가능성이 현저히 낮은 물량이 누적되어 향후 가용 주택공급이 체계적으로 과대추정될 우려가 있기 때문이다. 최근 국내 연구들 또한 주택공급이 인허가, 착공, 준공의 각 단계 사이에 상당한 시차를 수반하며, 사업기간의 장기화와 공정 지연이 사업 실현 가능성 및 공급 시점의 불확실성을 확대할 수 있음을 보여준다(Lee et al., 2025). 따라서 본 연구에서 장기 미준공 물량을 행정상 명시적 취소·철회·말소 건과 함께 비실현 공급으로 처리하여, 예측의 현실적 합성을 제고하고자 한다.

2. 종속변수

본 연구의 핵심 종속변수는 2020년 인허가 물량으로부터 발생한 실제 준공물량이다. 이를 y_m^{comp} 로 표기하며, 여기서 m 은 2020년 1월부터 최종 관측월까지의 각 월을 의미한다.

$$m = 1, 2, \dots, M \quad (\text{식 1})$$

월별 실제 준공물량은 2020년 인허가 건 가운데 해당 월에 준공일자가 확인된 물량을 합산하여 산출하였으며, 준공이 발생하지 않은 월의 값은 0으로 처리하였다. 예측 성능 평가는 2010~2019년 인허가~준공 데이터의 평균, 생존분석, Machine

Learning 학습을 한 각각의 모형이 2020년 인허가 물량에 대하여 산정한 월별 준공예상 물량과 실제 월별 준공물량 y_m^{comp} 를 비교하는 방식으로 수행한다. 즉, 본 연구는 개별 사업의 단순한 존속 여부가 아니라, 인허가 물량이 향후 어느 시점에 실제 준공으로 실현되는지를 예측하는 데 있다.

3. 설명변수

본 연구는 시장 변수, 금융 및 비용 변수, 공간 변수, 텍스트 변수와 같은 외생적 정보를 포함하지 않고, 공급 파이프라인과 일정 정보만을 설명 변수로 사용한다. 이는 세움터 행정자료만을 이용하여도 주택공급의 실현 경로를 상당 부분 설명할 수 있는지를 검토하고, 동시에 예측 시점에 실제로 관측 가능한 정보만을 활용함으로써 모형의 현실 적용 가능성을 높이기 위한 것이다. 설명변수는 네 범주로 구성된다(〈표 1〉).

첫째, 단계별 공급총량 변수이다. 여기에는 인

허가 물량, 착공 물량, 준공 물량 등 공급 파이프라인의 각 단계에서 관측되는 총량 정보가 포함된다. 둘째, 단계별 잔량(backlog) 변수이다. 여기에는 인허가 후 미착공 잔량, 착공 후 미준공 잔량 등 각 단계에 체류하고 있는 누적 물량이 포함된다. 셋째, 경과시간 변수이다. 여기에는 인허가 후 경과월수, 착공 후 경과월수, 미착공 지속기간, 미준공 지속기간 등 개별 사업별로 특정 단계에 머문 시간을 나타내는 정보가 포함된다. 넷째, 최근 흐름 변수이다. 여기에는 최근 1개월, 3개월, 6개월 동안의 누적 인허가·착공·준공 물량과 같이 단기적인 공급 흐름의 변화를 반영하는 지표가 포함된다. 이를 벡터 형태로 정리하면, 특정 월 m 의 설명변수는 X_m 으로 나타낼 수 있으며, 이는 단계별 공급총량 변수 X_m^{flow} , 단계별 잔량 변수 $X_m^{backlog}$, 경과시간 변수 $X_m^{duration}$, 최근 흐름 변수 X_m^{recent} 의 결합으로 표현된다. 따라서 본 연구의 예측 모형은 개념적으로 월별 실제 준공물량이 설명변수 벡터 X_m 의 함수로 결정된다고 볼 수 있으며, 예측값은 $\widehat{y_m^{comp}} = f(X_m)$ 와 같이 나타낼 수 있다.

이때 모든 설명변수는 예측 기준시점 이전까지 관측 가능한 정보만을 이용하여 구성한다. 따라서 특정 월의 준공물량을 예측할 때에는 해당 월 이후에만 확인 가능한 정보가 설명변수에 포함되지 않도록 하여, 미래 정보 누수 가능성을 차단한다. 이러한 변수 설계는 공급 실현량이 단순히 당기 인허가 물량의 함수가 아니라, 공급 파이프라인 내부에 누적된 물량이 어느 단계에 얼마나 오래 체류하고 있는가에 의해 결정된다는 가정에 기초한다. 즉, 동일한 인허가 물량이라 하더라도 착공 여부, 미착공 잔량의 규모, 착공 후 경과기간,

〈표 1〉 변수 구성 및 예측 기준시점별 가용 여부

변수명	변수정의	산출단위
인허가 물량	해당 월 인허가 물량	호
지역변수	지역 코드	범주형
사업유형	아파트	범주형
인허가-착공 소요기간	소요일	일
착공-준공소요기간	소요일	일
착공 후 미준공 잔량	잔여물량	호
기간별 인허가 누적	기준시점 이전 흐름	연속형
기간별 착공 누적	기준시점 이전 흐름	연속형
기간별 준공 누적	기준시점 이전 흐름	연속형

최근의 공급 흐름 등에 따라 실제 준공으로 이어지는 속도와 규모는 달라질 수 있다.

이와 같은 접근은 개별 사업의 단계별 이행구조와 지속기간 정보가 최종 공급 실현을 설명하는데 중요하다는 문제의식과도 맥락을 같이한다.

예측모형의 해석을 위해서는 각 설명변수가 실제 예측 기준시점에서 관측 가능한지 여부를 명확히 구분할 필요가 있다. 본 연구는 이 점을 반영하여 변수를 ‘기준시점 이전에 확정적으로 관측 가능한 변수’와 ‘사업 진행상황이 일부 반영된 조건부 변수’로 구분하였다. 특히 착공 후 경과월수, 착공 후 미준공 잔량, 최근 착공·준공 흐름 변수는 인허가 시점에서 직접 관측 가능한 정보가 아니므로, 이를 포함한 예측은 ‘인허가 시점의 순수 사전 예측’이라기보다 ‘사업 진행정보를 일부 반영한 조건부 예측’으로 해석되어야 한다. 따라서 본 연구는 예측 결과를 해석할 때 예측 기준시점과 정보 집합의 차이를 명시한다.

4. 비교모형과 제안모형

1) 단순 평균법

단순 평균법은 2010~2019년 인허가 후 준공이 완료된 수도권 아파트를 기준으로 평균 준공소요일을 산정하고, 이를 2020년 인허가 단지에 일괄 적용하는 방식이다. 즉, 과거 학습표본의 평균 인허가-준공 소요일을 PTC라 할 때, 개별 단지 i 의 예측 준공일은 실제 인허가일에 PTC를 더하여 산정한다. 연구에서 적용한 단순 평균법은 (식 2) 및 (식 3)과 같다.

$$\overline{PTC} = \frac{1}{N} \sum_{i=1}^N PTC_i \quad (\text{식 2})$$

$$\widehat{CompletionDate}_i = PermitDate_i + \overline{PTC} \quad (\text{식 3})$$

2) 생존분석 기반 준공예측 모형

본 연구의 생존분석 기반 예측모형은 지규현 외(2019)가 제시한 입주예정물량 추정 방법론의 기본 논리를 차용하여 구성하였다. 해당 연구는 공동주택 공급과정에서 관측시점 현재 아직 다음 단계로 이행하지 않은 잔여물량을 기준으로 단계별 발생률을 산정하고, 이를 이용하여 장래 공급물량을 추정하는 방식을 제안하였다. 특히 이미 일부가 진행되었으나 아직 사건이 발생하지 않은 물량과 분석기간 종료 시점까지 사건이 관측되지 않은 물량을 함께 반영할 수 있다는 점에서, 단순 평균적 배분방식보다 공급시차를 보다 현실적으로 반영할 수 있다는 장점을 가진다.

지규현 외(2019)의 원모형은 인허가, 모집공고, 입주 등 복수 단계의 전이과정을 대상으로 구성되어 있으나, 본 연구에서는 연구목적을 명확히 하기 위하여 이를 인허가 이후 준공까지의 소요기간 예측에 적용하였다. 즉, 본 연구에서 생존분석은 인허가 시점을 기준으로 시간의 경과에 따라 특정 월에 준공이 발생할 조건부 확률, 다시 말해 해당 시점까지 준공되지 않고 남아 있는 잔여물량 가운데 그 달에 실제로 준공되는 비율을 추정하는 방식으로 활용된다. 이는 선행연구가 강조한 “시점별 잔여물량 대비 발생률” 개념을 준공 예측에 맞게 단순화한 것이다.

이를 위해 먼저 인허가 시점이 동일한 주택물량을 하나의 동일 집단(cohort)로 정의한다. 특정 집단 n 에 대하여 인허가 직후의 총 물량을 P_n 이라 하고, 인허가 후 t 개월이 경과한 시점의 월초 기준 미준공 잔여물량을 U_t^m 이라 정의하면 다음과 같이 나타낼 수 있다.

$$U_t^m = P_n - \sum_{p=0}^{t-1} C_p^m \quad (\text{식 4})$$

여기서 C_p^m 은 동일 집단 n 에서 인허가 후 p 개월 시점에 실제로 준공된 물량이다. 그 외의 미준공 물량은 관측 종료 시점까지 준공이 발생하지 않은 진행 중 물량, 즉 중도절단 자료로 간주한다. 이는 최소물량을 처음부터 일괄 제외할 경우 초기 잔여물량이 과소계상되어 전이속도가 과대추정될 수 있다는 지규현 외(2019)의 논리를 반영한 것이다.

이와 같이 산정된 잔여물량을 바탕으로, 동일 허가건 n 의 인허가 후 t 개월 시점 준공 전이확률은 다음과 같이 정의한다.

$$k_t^n = \frac{C_t^m}{U_t^m} \quad (\text{식 5})$$

이는 인허가 후 t 개월이 되는 시점에 아직 준공되지 않은 물량 가운데 실제로 그 달에 준공된 비율을 의미한다. 이후 전체 학습기간에서 관측된 허가건의 전이확률을 평균하여 경과월별 대표 준공 스케줄 k_t 를 산출한다.

$$k_t = \frac{1}{N_t} \sum_{n=1}^{N_t} k_t^n \quad (\text{식 6})$$

여기서 N_t 는 인허가 후 t 개월 시점의 전이확률 계산이 가능한 허가건을 의미한다. 결국 k_t 는 “인허가 후 몇 개월이 경과하였을 때 평균적으로 어느 정도의 준공이 발생하는가”를 나타내는 경험적 준공 스케줄이라 할 수 있다. 선행연구가 모집공고 및 입주 단계에 대하여 동일한 방식의 전이 스케줄을 구성한 것과 마찬가지로, 본 연구는 이를 준공단계 예측에 맞추어 적용하였다. 학습자료로부터 추정된 대표 전이확률 k_t 를 이용하여 2020년 인허가건의 월별 준공물량을 예측할 수 있다. 특정 허가건 n 의 인허가 후 t 개월 시점 예측 준공물량은 해당 시점의 예측 잔여물량에 전이확률을 곱하여 다음과 같이 구한다.

$$\widehat{C}_t^m = k_t \times \widehat{U}_t^m \quad (\text{식 7})$$

그리고 다음 시점의 예측 잔여물량은 직전 잔여물량에서 예측 준공물량을 차감하는 방식으로 갱신한다.

$$\widehat{U}_{t+1}^m = \widehat{U}_t^m - \widehat{C}_t^m \quad (\text{식 8})$$

이 과정을 각각의 허가건에 대해 반복하면, 2020년 인허가 물량이 향후 어느 월에 얼마만큼 준공될 것인지를 월별로 배분할 수 있다. 최종적으로 특정 월 m 의 전체 예측 준공물량은 해당 월에 도달한 모든 허가건의 예측 준공물량을 합산하여 산출한다.

$$y_m^{\widehat{SURV}} = \sum_n \widehat{C}_{m,n} \quad (\text{식 9})$$

3) 인공지능 및 기계학습 기반 모형

본 연구의 인공지능 및 기계학습 기반 예측모형을 구축하기에 앞서, 수도권 아파트 인허가 단지의 준공소요일 예측에 적합한 알고리즘을 선정하기 위하여 ① Random Forest, ② XGBoost, ③ HistGradientBoosting, ④ LightGBM의 네 가지 알고리즘을 비교·검토하였다. 알고리즘의 비교는 단순한 예측정확도³⁾뿐 아니라 결측값 처리의 용이성, 범주형 변수 활용 가능성, 대용량 자료의 학습효율, 과적합 통제 가능성, 후속 갱신 및 운영의 실용성 등을 종합적으로 고려하는 방식으로 수행하였다.

〈표 2〉의 검토 결과 LightGBM은 예측정확도와 계산효율, 자료구조 적합성 측면에서 본 연구의 목적에 가장 부합하는 알고리즘으로 판단되었다. LightGBM을 최종 방법론으로 채택한 이유는 단순한 예측정확도뿐 아니라, 대용량 자료의 학습 효율성, 범주형 변수 처리의 실무적 편의성, 결측치 대응의 유연성, 과적합 통제를 위한 설정 가능성, 향후 갱신 및 운영 측면의 실용성을 종합적으로 고려했기 때문이다. 다만 이러한 선택은 특정 알고리즘의 일반적 우월성을 의미하지 않으며, 자료 특성과 운영환경에 따른 실무적 선택임을 전제로 한다.

LightGBM은 히스토그램 기반 구간화(histogram-based binning)와 효율적인 트리

〈표 2〉 모델별 비교검토 결과

모델	셀 정확도 ^a (%)	단지별 MAE ^b (일)	60일 이내 (%)	90일 이내 (%)
LightGBM 앙상블 ^c	71.2	28.6	87.1	94.7
HistGradientBoosting	25.4	93	53.9	63.8
Random Forest	20.3	102	49.2	59.8
XGBoost	25.4	101	49.4	62.1

주 : 1) ^{a)} 특정 연도-월 조합(‘셀’)의 예측 준공 세대수가 실제 대비 $\pm 15\%$ 이내인 셀의 비율. 본 모델의 핵심 목표 지표이며 목표 값은 85% 이상이다(예: 2023년 3월 실제 2,960세대 → 예측 2,680세대, 오차율 9.5% → 통과).

2) ^{b)} 평균 절대 오차. 단지별 예측 PTC 2(potential to commit 2)와 실제 PTC 차이의 절댓값 평균. 예) MAE=28.6일이면, 평균적으로 실제 준공일과 약 한 달 이내 오차로 예측하고 있음을 의미한다.

3) ^{c)} 본 연구에서 채택한 모델.

4) MAE, mean absolute error; LightGBM, light gradient boosting model; XGBoost, eXtreme gradient boosting.

분할 구조를 사용하여 대규모 구조화 자료에서 높은 학습 효율과 메모리 절감 효과를 보이며, 범주형 변수와 결측값을 비교적 유연하게 처리할 수 있다는 장점을 가진다. 따라서 본 연구는 최종 Machine Learning 기반 제안모형으로 LightGBM 4-way 앙상블을 채택하였다.

또한 LightGBM은 다음과 같은 장점을 가진다. 첫째, 수도권 아파트 인허가 자료는 세대수, 지역, 사업주체, 예정일자 관련 변수 등 수치형·범주형 변수가 함께 존재하고 일부 변수에는 결측이 포함되어 있는데, LightGBM은 히스토그램 기반 구간화와 트리 분할 구조를 통해 이러한 자료를 비교적 안정적으로 처리할 수 있다는 장점이 있다. 둘째, 대용량 구조화 자료에서 학습속도가

3) 예측정확도는 2010~2019년 학습데이터 중 10%의 랜덤 테스트 데이터를 설정, 90%의 나머지 학습데이터로 예측한 테스트 데이터의 정확도 여부로 판단하였다.

빠르고 메모리 효율성이 높아 반복적인 하이퍼파라미터 조정과 앙상블 구성에 유리하다. 셋째, 조기종료(early stopping), 정규화, 최소 자식 노드 제약 등 과적합 방지 장치를 함께 활용할 수 있어 실무형 예측모형으로 적합하다(Ke et al., 2017). XGBoost는 많은 선행연구에서 활용해 왔으며, 예측력 측면에서 유력한 대안이지만 결측 및 범주형 변수 처리의 부담이 상대적으로 크고, HistGradientBoosting은 구현이 간결하나 본 연구의 검증 기준에서 월별 정확도가 충분히 높지 않았으며, Random Forest는 해석 용이성과 강건성은 있으나 시점 예측의 정밀도 측면에서 상대적으로 낮은 성능이 나타났다. 이와 같이 네 가지 알고리즘의 특성과 예비 검증결과를 종합한 결과 본 연구에서의 최종 Machine Learning 기반 제안모형으로 LightGBM 앙상블을 채택하였다. 앙상블⁴⁾ 모델은 단일 모형이 특정 패턴에 편향되거나 특정 연도·유형의 표본에 과적합될 가능성을 줄이고, 준공소요일 분포의 폭넓은 패턴을 보다 안정적으로 반영하기 위한 것이다. 하위모형은 학습률, 리프 수, 최대 깊이, 최소 샘플 수 등을 달리 설정하여 균형형, 고복잡형, 경량형, 빠른 수렴

형의 역할을 수행하도록 설계하였다. 이와 같은 다중 모형 구조는 복수의 약한 학습을 결합하여 일반화 성능을 높이는 앙상블 접근의 장점을 활용하기 위한 것이다. 4-way 앙상블의 구성은 <표 3>과 같다.

개별 하위모형의 예측치는 단순 평균이 아니라 성능 가중 평균 방식으로 결합하였다. 즉, 학습 및 내부 검증 과정에서 상대적으로 오차가 작은 모형에 더 높은 가중치를 부여하고, 오차가 큰 모형에는 낮은 가중치를 부여하였다. 이러한 가중 앙상블은 특정 하위모형의 극단적 예측에 의해 전체 결과가 왜곡되는 문제를 완화하고, 개별 모형이 포착한 상이한 패턴을 최종 예측값에 반영하는 효과를 갖는다. 결과적으로 본 연구의 Machine Learning 모형은 단일 모델의 예측치가 아니라, 복수 LightGBM 모형이 생성한 예측을 성능 기반으로 통합한 합성 예측치라고 할 수 있다. 본 연구의 예측단위는 개별 인허가 단지이며, 종속변수는 각 단지의 PTC, 즉 인허가일로부터 실제 준공일까지의 소요일이다. Machine Learning 기반 모형은 (식 10)과 같이 고정된 시점이 아닌, 예측을 수행하는 특정 기준시점마다 새롭게 관측된 조

<표 3> 4-way 앙상블 구성

모델	학습률	트리 복잡도(리프)	최대 깊이	최소 샘플	특징
M1(균형형)	0.03	63	8	10	균형 잡힌 기본 모델, 안정적 성능
M2(고복잡)	0.02	127	10	8	세밀한 패턴 포착, 최대 복잡도
M3(경량형)	0.04	31	7	15	과적합 방지, 일반화 강조
M4(빠른 수렴)	0.05	31	6	20	빠른 수렴, 강건성 강조

4) 서로 다른 하이퍼 파라미터로 학습한 여러 모델의 예측 값을 결합하여 단일 모델보다 안정적인 결과를 도출하는 기법이다. 본 모델은 LightGBM 4개 모델을 학습 성능 기반 가중 평균으로 결합하였다.

건부 정보를 바탕으로 잔여 소요일을 추정한다.

$$PTC_{i,t} = f(X_{i,t}; \theta) \quad (\text{식 } 10)$$

여기서 $f(\cdot)$ 는 LightGBM 4-way 앙상블에 의해 학습된 비선형 예측함수이고, θ 는 모형의 하이퍼파라미터 및 추정된 트리 구조를 포괄하는 파라미터 집합이다. 이후 각 단지의 예측 준공시점은 실제 인허가일 $PermitDate_i$ 에 예측 소요일 PTC_i 를 더하여 산정한다.

$$CompletionDate_i = PermitDate_i + PTC_i \quad (\text{식 } 11)$$

그다음 동일한 연·월로 예측된 단지들의 세대수 $Units_i$ 를 합산하면, 특정 월 m 의 Machine Learning 기반의 예측 준공물량 y_m^{AI} 를 얻을 수 있다.

$$y_m^{AI} = \sum_{i \in m} Units_i \quad (\text{식 } 12)$$

본 연구의 Machine Learning 기반 제안모형은 단일 예측경로에만 의존하지 않고, 직접 예측과 2단계 분리 예측을 함께 활용하는 3단계 하이프라인으로 설계하였다. 첫째, 직접 예측 단계에서는 인허가 시점에 관측 가능한 단지 특성과 일정변수를 입력하여 인허가일부터 준공일까지의 총 소요일수 PTC를 한 번에 예측한다. 둘째, 2단계 분리 예측 단계에서는 인허가일부터 착공일까지

지의 기간 PTS와 착공일부터 준공일까지의 기간 STC를 각각 예측한 다음, 이를 합산하여 PTC를 재구성한다. 셋째, 최종 단계에서는 직접 예측 결과와 2단계 분리 예측 결과를 일정 비율로 혼합(blending)하여 최종 예측값을 산출한다. 이러한 구조를 채택한 이유는 총 소요일을 직접 예측하는 방식이 전체 기간의 전반적 분포를 안정적으로 포착하는 데 유리한 반면, 인허가에서 착공까지의 행정·사업 준비기간과 착공 이후 실제 공사기간은 결정요인과 변동성이 상이할 수 있기 때문이다. 따라서 공급지연의 원인을 보다 세분화하여 반영하려면 전체 기간을 두 부분으로 나누어 예측하는 방식이 보완적으로 필요하다고 판단하였다. 본 연구는 두 접근법의 장점을 결합하기 위하여 검증자료에서 최적의 혼합비율을 탐색하고, 그 결과를 최종 앙상블에 반영하였다(〈표 4〉).

IV. 분석 결과

1. 기술통계와 자료 특성

〈표 5〉는 2010~2019년 인허가건(학습데이터) 및 2020년 인허가건(테스트데이터)의 기술통계량이다.

분석 결과, 학습데이터는 5,091건 1,986,165

5) 본 연구는 2020년 1월부터 2020년 12월까지 인허가된 아파트를 독립적인 검증표본으로 설정하였다. 비록 2020년이 코로나19 팬데믹이라는 거시적 특수성이 존재했던 시기이나, 이를 검증표본으로 채택한 이유는 주택건설 사업의 장기적 특성 때문이다. 학습데이터 기준 인허가에서 준공까지는 평균적으로 1,011일이 소요되며, 사업 지연을 고려할 때 5년 안팎의 시차가 발생한다. 따라서 2021년 이후의 인허가 물량을 검증표본으로 설정할 경우, 분석 시점 현재 대다수의 사업이 아직 진행 중인 상태로 남아있게 되어 모형의 '최종 공급 실현 경로'를 온전히 평가하기 어렵다. 즉, 현재 시점에서 실제 준공 실적이 충분히 누적되어 외표본 검증에 활용할 수 있는 가장 현실적인 최신 자료가 2020년 인허가 데이터라고 판단하였다.

〈표 4〉 앙상블 개별 모델 성능 결과

모델	학습 MAE(일)	검증 MAE(일)	60일 이내(%)	90일 이내(%)	가중치(%)
M1: 균형형(LR=0.03, 63리프, 1,000트리)	60.5	28.6	86.0	96.6	26.5
M2: 고복잡(LR=0.02, 127리프, 1,000트리)	53.7	29.5	83.8	95.7	29.9
M3: 경량형(LR=0.04, 31리프, 800트리)	72.3	28.6	81.5	97.6	22.2
M4: 빠른 수렴(LR=0.05, 31리프, 600트리)	75.3	28.1	84.9	94.7	21.3
가중 앙상블(최종)		28.6	87.1	94.7	100

주 : 1) 학습 MAE, 학습셋 내 fit 품질 지표 (낮을수록 높은 가중치 부여); 검증 MAE, 내부 검증셋 기준 단지별 오차.
 2) M2가 학습 MAE 최저(53.7일)로 최고 가중치(29.9%) 부여. M3는 90일 이내 정확도 최고(97.6%).
 3) MAE, mean absolute error.

〈표 5〉 기초데이터 기술통계

구분	동수	세대수	비중(%)
학습 데이터 (2010~2019년)	5,091	1,986,165	100.00
검증 데이터 (2020년)	532	210,678	10.45

세대 수준이며, 검증데이터는 532건 210,678세대로 나타났다. 검증데이터의 규모는 학습데이터의 약 10.45% 수준으로 유사한 연구에서 활용하는 10%~20% 범위와 일치한다.⁶⁾

〈표 6〉은 학습데이터 기간 동안의 평균 인허가-준공 소요일자이다. 인허가에서 착공까지는 239일 약 8개월가량 소요되었으며, 착공에서 준공은 약 26개월, 인허가에서 준공은 33.7개월가량 소요되는 것으로 나타났다.

〈표 7〉은 학습데이터를 활용하여 산정된 생존 분석 준공확률 및 누적 준공확률이다.

〈표 6〉 평균 소요기간(단위: 일)

구분	소요기간
인허가-착공	239
착공-준공	771
인허가-준공	1,011

경과일은 인허가 이후 5개월부터 약 178개월까지로 나타났으며, 약 25% 수준이 관측종료 시점까지 미준공 상태로 남아 있음을 보여준다. 이러한 결과는 준공까지 이행이 단순 평균 수준을 넘어 이행되는 상당한 물량이 존재한다는 점이며, 단순 평균 접근만으로는 개별 프로젝트의 이후 준공경로를 충분히 설명하기 어렵다는 점을 시사한다.

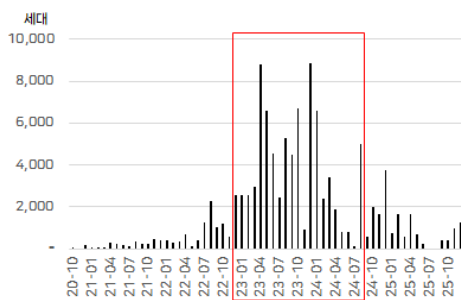
〈그림 1〉은 2020년 인허가 건으로부터 발생한 월별 실제 준공물량의 추이를 나타낸다. 준공은 2020년 10월부터 2025년 12월까지 확인되었으

6) 예측모형의 성능 평가는 모형 추정에 사용하지 않은 자료를 별도의 보류 검증자료(hold-out test set)로 설정하여 수행하는 방식이 일반적으로 활용된다. 예측 정확도는 모형 추정에 사용되지 않은 새로운 자료에 대한 성능으로 평가해야 하며, 테스트데이터의 규모는 전체 표본의 약 20% 수준이 일반적이나 표본의 길이와 예측기간에 따라 달라질 수 있다고 설명하였다(Hyndman and Athanasopoulos, 2021). 또한 Joseph(2022)은 학습·테스트 분할비율에 대해 일률적인 기준이 존재하지 않으며, 자료 및 모형 특성에 따라 적정 비율이 달라질 수 있음을 제시하였다. 한편 Bichri et al.(2024)은 60:40, 70:30, 80:20, 90:10의 학습·테스트 분할비율을 비교 분석하여 90:10 분할도 모형 성능 평가에 활용 가능한 실험 설계임을 보였다. 따라서 본 연구의 테스트데이터 규모는 외표본 검증을 위한 실무적·방법론적 범위 내에 있다고 볼 수 있다.

〈표 7〉 생존분석 준공확률 분석결과(단위:%)

경과월	누적 미준공율	경과월의준공 확률	누적 준공확률
1	100.00	0.00	0.00
12	99.13	0.14	0.87
24	96.34	0.44	3.66
36	74.32	1.41	25.68
48	59.31	0.90	40.69
60	52.91	0.42	47.09
72	47.66	0.21	52.34
84	42.24	0.31	57.76
96	36.94	0.82	63.06
108	33.47	0.00	66.53
120	31.29	0.31	68.71
132	29.99	0.00	70.01
144	28.75	0.00	71.25
156	27.79	0.04	72.21
168	25.98	0.00	74.02
178	25.62	0.00	74.38

주 : 경과월 전체는 생략, 연도에 해당하는 누적 준공확률임.



주 : 붉은색 박스 부분이 실제 준공물량이 집중된 구간으로 2023년 1월~2024년 7월까지임.

〈그림 1〉 2020년 인허가데이터 실제 준공 물량 추이

며, 총물량은 약 10.8만 호 수준으로 20년 1년간 인허가 물량 중 약 52% 수준이다. 준공물량은

2023년 1월~2024년 7월에 약 68%가 집중되는 패턴을 보였다. 이러한 패턴은 공급실현량이 단순 추세가 아니라 파이프라인 내 미준공 발생사유와 경과시간 구조에 의해 결정된다는 점을 보여준다.

2. 월별 준공물량 예측성능 비교

〈표 8〉은 단순평균법, 생존분석 모형, Machine Learning 기반 모형의 월별 준공물량 예측성능을 비교한 결과이다. 예측성능은 각 월별 실제값과 추정값의 차이 절대 오차를 활용하였다. Machine Learning 기반 모형의 월별 평균 절대오차는 166, 단순평균법은 12,265 생존분석 모형은 952로 Machine Learning 기반 모형의 절대오차 평균이 가장 낮았다. 이는 Machine Learning 기반 모형이 단계별 잔량과 경과기간 간의 비선형 상호작용을 여타의 예측방법보다 효과적으로 포착하였기 때문으로 해석된다. 단순평균법은 특정 월에 준공이 집중되거나 지연이 누적되는 구간을 정확하게 포착하지 못했다. 생존분석 모형은 개별 프로젝트의 준공확률을 반영한다는 점에서 평균법보다 우수했지만, 월별 총량의 급격한 변동을 설명하는 데는 일정한 한계를 보였다. 반면 Machine Learning 기반 모형은 월별 분포의 봉우리와 지연 해소 구간을 상대적으로 잘 추적하여, 시점별 공급실현 패턴을 보다 정밀하게 예측하였다.

3. 누적 준공물량 비교

월별 정확도와 함께 누적 준공물량을 비교한

〈표 8〉 실제준공량과 평균법 기반예측, 생존분석예측, Machine Learning 예측 값 및 절대오차

준공연월	실제준공량	평균법예측	생존분석예측	Machine Learning 예측	평균법_절대오차	생존분석_절대오차	Machine Learning_절대오차
20-06			8				
20-07			21				
20-08			25				
20-09			36				
20-10	21		53	26		32	5
20-11			69	24			
20-12	164		95	53		69	111
21-01	36		35	52		1	16
21-02	49		124	57		75	8
21-03	86		149	29		63	57
21-04	319		179	76		140	243
21-05	226		241	103		15	123
21-06	180		230	160		50	20
21-07	123		257	194		134	71
21-08	354		294	209		60	145
21-09	251		339	247		88	4
21-10	251		352	286		101	35
21-11	449		366	204		83	245
21-12	399		380	483		19	84
22-01	404		444	226		40	178
22-02	313		321	419		8	106
22-03	347		212	309		135	38
22-04	683		744	536		61	147
22-05	152		984	381		832	229
22-06	417		1,313	510		896	93
22-07	1,241		1,521	1,001		280	240
22-08	2,260		1,644	2,166		616	94
22-09	1,023		1,822	1,067		799	44
22-10	1,178	5,070	2,100	1,120	3,892	922	58
22-11	593	17,516	2,721	630	16,923	2,128	37
22-12	2,555	12,307	2,633	2,398	9,752	78	157
23-01	2,544	11,049	3,272	2,480	8,505	728	64
23-02	2,562	9,068	3,394	2,171	6,506	832	391
23-03	2,960	12,403	4,911	2,680	9,443	1,951	280
23-04	8,781	19,705	3,998	8,511	10,924	4,783	270

〈표 8〉 Continued

준공연월	실제준공량	평균법예측	생존분석예측	Machine Learning 예측	평균법_절대오차	생존분석_절대오차	Machine Learning_절대오차
23-05	6,577	11,893	4,110	6,211	5,316	2,467	366
23-06	4,534	11,372	3,985	4,447	6,838	549	87
23-07	2,437	16,223	4,102	2,405	13,786	1,665	32
23-08	5,291	12,852	3,745	5,453	7,561	1,546	162
23-09	4,517	38,138	3,825	3,849	33,621	692	668
23-10	6,701	33,082	3,828	6,813	26,381	2,873	112
23-11	948		3,467	983		2,519	35
23-12	8,835		2,748	8,312		6,087	523
24-01	6,605		2,956	7,235		3,649	630
24-02	2,418		2,213	2,571		205	153
24-03	3,435		2,491	2,701		944	734
24-04	1,895		2,253	1,921		358	26
24-05	833		2,046	1,045		1,213	212
24-06	786		2,077	910		1,291	124
24-07	125		1,875	282		1,750	157
24-08	4,998		1,611	4,662		3,387	336
24-09	606		1,993	466		1,387	140
24-10	1,975		1,169	2,055		806	80
24-11	1,658		1,419	1,508		239	150
24-12	3,739		1,356	3,837		2,383	98
25-01	748		1,154	649		406	99
25-02	1,644		1,028	1,263		616	381
25-03	592		1,150	609		558	17
25-04	1,674		1,219	1,563		455	111
25-05	697		831	427		134	270
25-06	242		987	407		745	165
25-07			629				
25-08			752				
25-09	418		123	392		295	26
25-10	426		472			46	
25-11	1,003		283	770		720	233
25-12	1,273		142	1,320		1,131	47
합계	108,551	210,678	97,326	103,874	159,448	57,135	9,767
평균	1,809	16,206	1,453	1,731	12,265	952	166

결과도 중요하다. <표 9>는 모형별 누적 준공물량 예측치와 실제 누적 실적을 비교한 결과이다. 평균법의 경우 취소, 지연 등을 전혀 고려하지 못해 누적합계 비중이 실제 준공치의 약 194% 수준으로 크게 나타났으며, 생존분석은 약 89.6%, Machine Learning 예측은 약 95.6%의 누적물량 정확도가 나타났다.

Machine Learning 기반 모형은 월별 준공물량 오차뿐 아니라 누적 총량 기준으로도 실제 준공경로를 가장 안정적으로 추적하였다. 이는 Machine Learning 모형의 장점이 단순히 단기 타이밍을 잘 맞추는 데 그치지 않고, 최종 공급실현 총량의 경로를 보다 정확히 재현하는 데에도 있음을 의미한다.

4. 소결

공급이 중요한 주택시장에서 “현재 인허가, 착공, 준공이 몇 호인가”보다 “허가된 물량이 얼마나 오랫동안 미착공 상태로 남아 있는가”와 “착공된 물량이 평균적으로 어느 정도 진행되었는가”가 향후 공급전망에 더 직접적인 정보를 제공한다. 따라서 공급 모니터링 체계는 단순 실적 집계에서 벗어나 단계별 잔량과 체류시간을 함께 관리

하는 방향으로 발전할 필요가 있다. 본 연구의 결과는 주택공급 파이프라인에서 나타나는 비선형적 이행 패턴이 평균 기반 접근이나 단순한 선형 구조만으로는 충분히 설명되기 어렵다는 점을 시사한다. 따라서 Machine Learning 기반 모형은 예측정확도 향상뿐 아니라, 공급지연의 병목이 어느 단계에서 발생하는지를 해석하는 데도 유용하다고 판단된다.

종합하면, 2010~2019년 인허가 이력을 학습하여 2020년 인허가 건의 월별 준공실적을 예측한 결과, Machine Learning 기반 모형은 단순 평균법과 생존분석에 비해 전반적으로 낮은 예측 오차를 보였으며, 특히 월별 공급집중 구간과 누적 준공경로를 상대적으로 더 정확하게 추적하는 것으로 나타났다. 특히 일정 정보의 세분화는 예측력 향상에 중요한 역할을 하였으며, 공급총량 자체보다 단계별 잔량과 경과기간 구조가 향후 준공실현량의 핵심 설명요인임이 확인되었다.

한편 Machine Learning 기반 모형의 월별 예측은 초기 구간에서는 상대적으로 불안정한 양상을 보였으나, 경과기간이 충분히 누적된 이후 구간에서는 높은 적중률이 관찰되었다. 이는 허가 직후 초기 구간일수록 준공실현 여부의 불확실성이 크고, 일정 정보가 누적될수록 준공 패턴이 보다 명확해지기 때문으로 해석된다. 특히 공급실현이 집중되는 구간에서는 Machine Learning 모형이 실제 월별 준공분포를 안정적으로 재현하는 것으로 나타났다. Machine Learning 기반 제안모형은 수도권 아파트 인허가 단지의 준공 소요일과 월별 준공물량을 예측하는 데 있어 유의미한 수준의 설명력과 실무 적용 가능성을 보였다.

<표 9> 누적 준공물량 비교

구분	참값	평균법	생존분석	Machine Learning 예측
누적 합계	108,551	210,678	97,326	103,874
비중(%)	100	194.08	89.66	95.69

특히 공급 파이프라인과 일정 정보만을 활용했음에도 불구하고, 월별 수준과 누적 수준 모두에서 매우 안정적인 성능을 나타냈다는 점은 본 연구의 중요한 결과이다.

V. 결론

본 연구는 2010년 1월 1일부터 2019년 12월 31일까지 인허가된 주택사업을 학습자료로, 2020년 1월 1일부터 2020년 12월 31일까지 인허가된 주택사업을 테스트자료로 설정하여, 2020년 인허가 주택의 이후 기간별 준공실적을 예측하는 모형을 비교·평가하였다. 연구의 핵심은 “2020년에 허가된 주택사업이 이후 어느 시점에 얼마만큼 실제 준공되는가”를 예측하는 데 있었다. 이를 위해 단순평균법, 생존분석, Machine Learning 기반 모형을 동일한 외표본 평가를 안에서 비교하고, 공급 파이프라인과 일정 정보만으로 어느 정도의 예측력을 확보할 수 있는지를 검증하였다.

분석 결과는 다음과 같이 요약할 수 있다. 첫째, Machine Learning 기반 모형은 월별 준공물량 예측에서 단순평균법과 생존분석보다 낮은 예측 오차를 보였다. 이러한 성능 격차는 알고리즘 자체의 우수성뿐만 아니라, 인허가 시점의 고정된 정보(사전 예측)에 의존하는 비교 모형들과 달리 제안 모형이 착공 이후의 갱신된 진행 정보를 조건부로 활용하여 정보집합을 확장(nowcasting)한 데서 상당 부분 기인한다. 둘째, 월별 분포뿐 아니라 누적 준공물량 기준으로 Machine Learning 기반 모형은 실제 실적경로를 보다 정

확하게 추적하였다. 셋째, 단순 총량 정보나 평균 소요일에 의존하는 전통적 방식과 달리, 단계별 잔량과 경과기간 정보를 통합적으로 학습한 머신러닝 제안모형이 가장 우수한 예측 성능을 보였다. 이는 공급실현량 예측에 있어 파이프라인 내 물량의 체류상태와 진행속도라는 구조적 정보를 모형화하는 것이 정책적으로 매우 유효한 접근임을 시사한다.

본 연구의 학술적 의의는 세 가지이다. 첫째, 국내 주택시장 연구에서 상대적으로 부족했던 주택 공급 관련 연구 특히 준공경로 예측을 다루었다는 점이다. 둘째, 가격이나 거래가 아니라 공급실현량을 예측대상으로 설정함으로써, 부동산 Machine Learning 연구를 공급 파이프라인 문제로 확장하였다는 점이다. 셋째, 단순평균법과 생존분석을 기준모형으로 설정하고, Machine Learning 기반 모형의 추가 예측력을 외표본 자료를 통해 검증하였다는 점이다.

정책적 시사점은 다음과 같다. 첫째, 공급정책의 성과평가는 단순 인허가 실적에 머물러서는 안 되며, 허가 이후 실제 준공으로 이어지는 경로를 추적해야 한다. 둘째, 공급 모니터링 체계는 단순 실적 집계보다 미착공 잔량, 미준공 잔량, 인허가 후 경과월수, 착공 후 경과월수와 같은 일정기반 정보를 함께 관리할 필요가 있다. 셋째, 인허가 단계의 조기 진단을 넘어, 착공 이후 누적되는 행정 정보를 적극 활용한 단기 준공물량의 정밀한 전망 체계가 필요하다. 본 연구의 예측모형은 현재 파이프라인이 어느 단계에서 병목을 겪고 있는지 파악하고 조건부 공급 변동을 단기적으로 추적하는 실무적 기초도구로 활용될 수 있다.

다만 본 연구는 몇 가지 한계를 가진다. 첫째, 설명변수를 공급 파이프라인과 일정 정보로 제한하였기 때문에, 금리, 미분양, 건설공사비, PF 경색, 지역수요 변화와 같은 외부 환경요인을 직접 반영하지 못한다. 향후 연구에서는 본 연구가 검증한 행정 데이터기반 모형에 거시경제 및 시장 변수를 결합하여, 조건 변화에 따른 물량 변화를 시뮬레이션할 수 있는 확장 모형이 필요하다. 둘째, 현재까지 준공이 확인되지 않은 일부 2020년 허가건은 여전히 진행 중일 수 있으며, 이 경우 장기 경로를 완전히 관측하지 못했을 가능성이 있다. 셋째, 취소의 정의와 자료처리 방식에 따라 보조분석 결과는 일부 달라질 수 있다. 따라서 본 연구의 결과는 “공급행정자료와 일정 정보만으로 확보할 수 있는 기준 성능”으로 이해하는 것이 적절하다.

셋째, 본 연구는 단순평균법, 생존분석방법, 머신러닝 방법론의 예측 결과를 비교하였으나, 이들 방법은 본래의 목적과 산출구조가 동일하지 않다. 단순평균법은 고정적 요약치에 기반한 단순 예측이며, 생존분석방법은 사건발생의 시간구조 또는 확률을 설명하는 데 강점이 있고, 머신러닝은 예측오차 최소화에 초점을 둔다. 따라서 MAE와 같은 오차지표 비교는 실무적 예측력의 참고자료로는 유용하지만, 이를 근거로 방법론의 일반적 우열을 단정하는 데에는 한계가 있다. 특히 생존분석 기반 접근이 분포 또는 확률 구조를 다루는 반면, 본 연구의 비교 대상인 추정치는 점예측 형태로 변환되어 비교되었다. 이 경우 비교결과는 ‘실무적 점예측 성능’에 대한 제한적 비교로 해석되어야 하며, 확률예측의 적합성이나 분포 추

정의 타당성까지 포괄하는 평가는 아니다. 머신러닝모형의 오차가 더 낮게 나타났더라도, 이는 해당 모형이 비교 대상보다 본질적으로 우월하다는 의미보다는, 주어진 정보집합과 평가지표하에서 예측 중심 과업에 더 적합했음을 시사하는 결과로 해석할 필요가 있다.

넷째, 본 연구의 제안모형은 비교모형 대비 알고리즘(머신러닝)과 투입 정보(파이프라인 변수)가 모두 고도화된 형태이므로, 예측 성능의 향상이 알고리즘 자체의 우수성에서 기인한 것인지, 추가된 변수의 설명력에서 기인한 것인지를 분리하여 검증하지 못했다는 방법론적 한계가 있다. 향후 연구에서는 동일한 알고리즘 내에서 총량 변수만을 투입한 모형과 파이프라인 변수를 모두 투입한 모형을 비교하는 통제 실험(ablation study)을 통해 개별 변수군의 순수 기여도를 규명할 필요가 있다.

향후 연구에서는 본 연구의 일정기반 모형을 기준선으로 삼아, 시장 변수, 금융·비용 변수, 공간 변수, 텍스트 변수를 단계적으로 결합한 확장 모형을 검토할 필요가 있다. 구체적으로는 주택 가격 변동률, 거래량, 미분양 물량, 시장금리, 건설공사비, 접근성, 시공사 등의 정보를 추가할 수 있을 것으로 사료된다. 그러나 이러한 확장은 다음 단계의 과제로 남겨두고, 본 연구는 먼저 공급 파이프라인과 일정 정보 그 자체만으로도 유의미한 공급실현량 예측체계를 구성할 수 있음을 보였다는 점에 의의를 둔다.

ORCID

박태이 <https://orcid.org/0009-0003-9931-6012>

주승민 <https://orcid.org/0009-0006-2108-2625>

참고문헌

1. 김이환, 김형준, 류두진, 조훈. (2022). 기계학습 방법론을 활용한 아파트 매매가격지수 연구. *부동산 분석*, 8(3), 1-29.
2. 김형석, 이영성. (2008). 우리나라 신규주택공급의 장기 가격탄력성 연구: 1980년대 후반 이후의 신규 주택공급을 중심으로. *대한국토·도시계획학회 추계 학술대회 논문집 1A*(pp. 869-877). 대한국토·도시계획학회.
3. 박대현, 김정환, 류두진. (2021). 기계학습 기반 주택시장의 조기경보체계. *부동산분석*, 7(1), 29-45.
4. 이진원, 민병학, 김세용. (2018). 재개발사업 시행에 미치는 영향요인 분석: 생존분석의 모수분포를 가정한 분석모델을 중심으로. *국토지리학회지*, 52(1), 1-10.
5. 이선구. (2025). XGBoost 기반 부동산 자동 가치산정모형(AVM)의 실증 분석: 은평구 다세대 주택 실거래가를 중심으로. *부동산분석*, 11(2), 21-40.
6. 이준용, 손재영. (2013). 패널분석을 이용한 대도시 주택가격 추이 분석. *부동산학연구*, 19(4), 71-86.
7. 지규현, 곽하영, 이창무, 한계선. (2019). 생존분석 개념을 이용한 입주예정물량 추정 방법론 연구. *부동산연구*, 29(4), 69-83.
8. 지규현, 최성호, 주현태, 이창무. (2017). 수도권 주택건설 착공시기에 대한 생존분석. *주택연구*, 25(3), 117-132.
9. 홍정의. (2021). 랜덤 포레스트 알고리즘을 통한 주택 대량평가모형 연구. *부동산분석*, 7(1), 1-28.
10. Bichri, H., Chergui, A., & Hain, M. (2024). Investigating the impact of train / test split ratio on the performance of pre-trained models with custom datasets. *International Journal of Advanced Computer Science and Applications*, 15(2), 331-339.
11. Cai, C., Zou, Y., Peng, Y., & Zhang, J. (2012). smcure: An R-package for estimating semiparametric mixture cure models. *Computer Methods and Programs in Biomedicine*, 108(3), 1255-1260.
12. Chen, S., Jin, H., & Li, L. (2023). Analysis and comparison of house price prediction based on XGBoost and LightGBM. *Advances in Economics, Management and Political Sciences*, 46, 55-61.
13. Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2), 187-202.
14. Green, R. K., Malpezzi, S., & Mayo, S. K. (2005). Metropolitan-specific estimates of the price elasticity of supply of housing, and their sources. *American Economic Review*, 95(2), 334-339.
15. Hansen, P. R., & Timmermann, A. (2012). *Choice of sample split in out-of-sample forecast evaluation*. https://rady.ucsd.edu/_files/faculty-research/timmermann/samplesplitmining2012_feb07.pdf
16. Hosmer, D. W., Lemeshow, S., & May, S. (2008). *Applied survival analysis: Regression modeling of time-to-event data* (2nd ed.). John Wiley & Sons.
17. Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts.

18. Joseph, V. R. (2022). Optimal ratio for data splitting. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 15(4), 531–538.
19. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30* (pp. 3146–3154). NIPS.
20. Kleinbaum, D. G., & Klein, M. (2012). *Survival analysis: A self-learning text* (3rd ed.). Springer.
21. Lahmiri, S., Bekiros, S., & Avdoulas, C. (2023). A comparative assessment of machine learning methods for predicting housing prices using Bayesian optimization. *Decision Analytics Journal*, 6, 100166.
22. Lee, J. G., Jeong, S., & Choi, J. (2025). Time lag analysis of real estate supply policies for housing construction projects. *Journal of Asian Architecture and Building Engineering*, 24(4), 2788–2799.
23. Makridakis, S., & Hibon, M. (2000). The M3-competition: Results, conclusions and implications. *International Journal of Forecasting*, 16(4), 451–476.
24. Mayer, C. J., & Somerville, C. T. (2000). Residential construction: Using the urban growth model to estimate housing supply. *Journal of Urban Economics*, 48(1), 85–109.
25. Pita, R. P., de Carvalho, A. R., Barbosa, R. M., Cury, A. A., & Mendes, J. C. (2025). A systematic review of the use of machine learning in the prediction of house pricing. *Computational Economics*, 67(5), 3993–4032.
26. Saiz, A. (2010). The geographic determinants of housing supply. *The Quarterly Journal of Economics*, 125(3), 1253–1296.
27. Tapia, J., Chavez-Garzon, N., Pezoa, R., Suarez-Aldunate, P., & Pilleux, M. (2025). Comparing automated valuation models for real estate assessment in the Santiago Metropolitan Region: A study on machine learning algorithms and hedonic pricing with spatial adjustments. *PLOS ONE*, 20(3), e0318701.
28. Zhang, Y., & Miller, E. J. (2023). Predicting housing construction period based on a Cox proportional hazard model: An empirical study of housing completions in the Greater Toronto and Hamilton Area. *Environment and Planning B: Urban Analytics and City Science*, 50(6), 1624–1644.

논문 접수 일: 2026년 5월 8일

심사(수정)일: 2026년 7월 3일

게재 확정 일: 2026년 7월 10일

국문초록

본 연구는 수도권 아파트를 대상으로 2010년 1월부터 2019년 12월까지의 인허가 이력을 학습자료로, 2020년 인허가 물량을 테스트자료로 설정하여 허가 이후 실제 준공으로 이어지는 주택공급 실현량을 예측하는 모형을 평가하였다. 비교모형인 단순 평균법과 생존분석이 인허가 시점의 정보에 기반한 정태적 사전 예측(forecasting)인 반면, 제안모형인 LightGBM(light gradient boosting model) 4-way 앙상블 기반 Machine Learning 모형은 착공 후 경과월수 등 지속적으로 갱신되는 정보를 활용하는 동적 단기 전망(nowcasting)으로 정보집합의 확장을 시도 후 상호 비교 평가 하였다. 설명변수는 시장·금융·공간 변수 등을 제외하고 공급 파이프라인과 일정정보로 한정하였다. 분석 결과, 확장된 파이프라인 정보를 활용한 Machine Learning 기반 모형은 월별 준공물량 예측에서 단순 평균법과 생존분석보다 낮은 예측오차를 보였으며, 누적 준공물량 기준에서도 실제 공급경로를 가장 안정적으로 추적하였다. 이는 공급총량 자체보다 단계별 잔량과 경과기간 구조가 향후 준공실현량 전망의 핵심 정보임을 시사한다. 본 연구는 갱신되는 행정정보만으로도 실효성 있는 공급량 예측체계를 구축할 수 있음을 보였다는 점에서 의의를 가진다.

주제어 : 주택공급 실현량, 준공물량예측, 생존분석, Machine learning, LightGBM(light gradient boosting model) 앙상블