pISSN : 2465-9754 eISSN : 2508-1292 https://doi.org/10.30902/jrea.2025.11.1.109

Journal of Real Estate Analysis

http://www.ejrea.org *April 2025, Vol.11, No.1, pp. 109~133*

토지 감정평가서의 텍스트 마이닝 기반 공매 낙찰가 예측*

Text Mining Based Price Prediction in Public Auctions using Land Appraisal Reports

문혜정** · 조남욱***

HyeJung Moon · Nam-Wook Cho

Abstract

This study aimed to predict land prices of public auctions based on appraisal reports. Analyzing 24,047 land transactions from Onbid (2017-2024), the dataset included property details, bidding information, appraisal reports and regional data on population · households (2,784 cases), local taxes (2,080) and land prices (28,235). Text mining extracted qualitative variables, while instrumental variable and two-stage least squares investigated the causation of auction prices. Performances were assessed by adj.R^2, root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), successful bids and probability of successful bids. Predictions for the rounds of bidding and the prices of land categories indicates that MAPE performed better in quantitative analysis, whereas MAE and RMSE were superior in integrated analysis. The probability of successful bids was about 2.89% to 22.15% higher for integrated analysis cases on every case. Quantitative variables (e.g.,the ratio of the prices on public auctions in the neighborhood) were effective with ample historical data and stable economic conditions. In contrast, qualitative factors (e.g., transportation, sentiment and legal issues) were valuable either data were limited or economic conditions were unstable. This study pioneered how to determine qualitative variables by analyzing appraisal reports and identified the characteristics of quantitative and qualitative factors to propose optimal situation-specific predictions. Additionally, this study contributes to public auctions and general valuation of lands by deriving price factors with high explanation but low prediction errors.

Keywords: Public auction, Appraisal report, Price prediction, Text mining, PropTech

* 본 연구는 2025년도 정부(산업통상자원부)의 재원으로 한국산업기술진흥회의 지원을 받아 수행되었음(P0017123, 2025년 산업혁신인재성장지원사업).

** 서울과학기술대학교 IT정책전문대학원 박사과정(주저자) | Ph.D. Program, Graduate School of Public Policy and Information Technology, Seoul National University of Science & Technology | First Author | hyejung.moon@gmail.com |

*** 서울과학기술대학교 산업공학과 교수(교신저자) | Professor, Department of Industrial Engineering, Seoul National University of Science & Technology | Corresponding Author | nwcho@seoultech.ac.kr |

Ⅰ. 서론

본 연구는 토지의 고유한 특성과 활용가능성에 초점을 맞춘다. 토지가격에는 거래 당시의 부동산 경기나 대출 규제 정책 등 외부 요인보다 토지 자체의 고유한 특성과 그 활용 가능성이 더 큰 영향을 미칠 수 있다. 토지의 가치분석을 위해 본 연구는 자료의 수집이 용이한 공공데이터로 개방된 공매물건의 매각정보와 감정평가서 내용을 기반으로 낙찰가격을 분석한다. 공매자료는 채권정보, 배분요청내용 및 매각이력이 모두 공개되어 있어 정보의 투명성과 포괄성 측면에서 토지의 가치 판단을 위해 적합한 자료이다.

대부분의 공매는 조세채권 회수를 목적으로 진행되며, 감정평가서는 토지의 질적 특성과 고유정보를 문서 형태로 포함하고 있다. 특히 주변 환경 정보나 시장 상황과는 별개로 토지 자체의 특성과 감정평가서의 종합 의견을 분석할 경우 토지의 활용 가치를 구체적으로 파악할 수 있지만, 비정형 문장을 통계적 방법으로는 분석하기는 어렵다. 최근 생성형 AI 기술의 발달은 이러한 비정형 데이터를 효과적으로 분석할 수 있는 새로운 가능성을 제시한다. 이에 본 연구는 텍스트 마이닝 기법을 활용하여 감정평가서의 문장 정보를 분석하고, 도출된 맥락 변수를 기반으로 공매 토지의 낙찰가를 예측하여 토지의 고유가치를 기반으로 활용 가능성을 찾는 것에 중점을 둔다.

공매 정보를 체계적으로 분석하고 이를 통해 낙찰가를 예측하는 모델이 개발된다면, 이전보다 낮은 위험과 가격으로 부동산을 매수할 가능성이 커질 것이다. 또한, 경·공매 물건이 적정 가격으로 매각될 경우, 국가 등 채권자들이 자산에 대한 권리와 처분을 효율적으로 행사할 수 있다. 이러한 결과는 부동산 시장의 효율성을 높이고, 경·공매 절차를 통해 다양한 이해관계자의 이익도 증진시키리라 기대한다.

본 연구의 목적은 2017년부터 2024년까지 공매로 매각된 약 2.4만 건의 토지 낙찰 데이터를 분석하여, 낙찰가격의 형성 요인을 밝히고 예측 모델을 구축하는 것이다. 기존의 연구들이 주로 정량 변수에 기반한 분석에 집중해온 것과 달리, 본 연구는 감정평가서에 포함된 정성적 정보를 텍스트 마이닝 기법으로 정제·추출하여 변수화하고, 이를 분석 모델에 포함하여 예측 성능을 비교 분석한다. 특히 감정평가서에서 도출된 맥락적 정보가 정량 변수와 어떤 차별성을 가지며 낙찰가 예측에 어떠한 기여를 할 수 있는지를 분석한다. 이 과정에서 입찰 회차별 및 지목별로 낙찰가격에 영향을 미치는 주요 변수의 상대적 영향력과 예측 성능의 차이를 살펴보고, 정성변수의 활용 가능성과 한계를 확인해보겠다.

Ⅱ. 선행연구 고찰

1. 한국의 경·공매 제도

부동산 경매는 통상 법원경매를 의미하며 강제경매와 임의경매가 있다. 채권자가 법원을 통해 채무자의 재산을 처분하여 채권을 회수한다는 점에서 동일하나, 근거·신청 주체·절차 등에서 차이가 있다. 공매는 국가나 지자체, 공공기관, 또는

 https://doi.org/10.30902/jrea.2025.11.1.109

민간이 보유한 재산을 공개적으로 매각하는 것을 의미하나, 통상 국세징수법에 따른 한국자산관리공사의 체납압류재산 처분 공매를 의미한다. 전통적인 경매이론 측면에서 볼 때, 경·공매 모두 형식면에서 입찰자의 가치 평가가 서로 연관되어 있으며, 자신의 정보외에도 타인의 신호도 중요한 역할을 하는 연계가치모형이다. 즉, 감정평가서와 매각관련 정보가 입찰 전 제공되는 신호로써 작용하며, 최고가격으로 낙찰이 결정되는 매각일등가격 밀봉입찰 방식이다. 따라서, 입찰자 간 기대 수익과 전략에 따라 결과의 다양성이 발생하며, 최저입찰가격은 공공 수익을 확보하면서 거래 활성화도 유도하는 전략 수단으로 작용한다(Milgrom, 1985).[1)]

부동산 가격을 분석하는 대표적인 접근 방식은 다양한 속성이 가격에 미치는 영향을 분해하여 기여도를 추정하는 헤도닉 가격 모형이다(Rosen, 1974). 그러나 경·공매처럼 입찰 방식에 따라 가격이 결정되는 경우 제도적 조건과 입찰자의 전략이 핵심이 되는 게임이론 기반 가격결정모형이 적합하다(Milgrom, 1985, 2019). 게임이론적 경매 이론 관점에서 볼 때, 부동산 경·공매는 입찰자 간 전략적 상호작용을 수리적으로 분석하고, 최적 입찰전략과 효율적인 자원 배분을 가능하게 한다(Milgrom, 1985, 2004, 2019). 이를 통해 매각자는 수익을 극대화하고, 입찰자는 정보에 따른 합리적 선택을 하며, 결과적으로 시장 기반의 실질적인 가치가 형성된다. 이러한 점에서 부동산 경·공매는 시장가격을 예측하기에 매우 적합한 분석 대상이라 할 수 있다. 경매 정보는 매각시점에만 공개되나 공매 정보는 매각 후에도 공공데이터로 제공되므로, 본 연구는 분석의 일관성과 접근 가능성을 고려하여 공매 정보를 연구 대상으로 선정하였다.

2. 선행 연구

감정평가서 분석을 통한 토지 공매의 낙찰가 예측과 일치하는 선행연구는 드물기 때문에 토지와 건물 대상 낙찰가 분석(또는 예측)과 자연어처리를 활용한 부동산 가격 분석으로 구분하여 검토하였다. 토지 관련 낙찰가격 관련 연구는 다음과 같다. Ooi et al.(2006)은 1990~2002년간 싱가포르의 경매 202건을 분석하여 면적, 도심거리, 용도지역, 개발이력, 입찰기업의 상장여부, 입찰자 수, 매매지수 등이 낙찰요인임을 밝혔다. Tse et al.(2011)은 홍콩에서 집행된 1993~2002년간 경매 223건의 낙찰가격요인이 경매의 불확실성,

1) 전통적인 경매 이론은 경매의 유형 기준과 경매의 경제적인 공통 특성으로 구성된다. 유형 기준은 가치 구조, 정보의 비대칭성과 정보 구조, 경매 형식 세 가지이다. 공통 특성은 전략적 행동, 최저입찰가격 두 가지이다. 가치 구조는 입찰자가 개인적인 판단에 기반해 가치를 결정하는 사적 가치 모형, 물건의 객관적 가치는 동일하나 입찰자 간 추정이 다른 공통 가치 모형, 입찰자들의 가치 평가가 서로 연관되는 연계 가치 모형으로 구분된다. 정보 구조는 공개입찰과 비공개입찰, 입찰 전 제공되는 신호의 존재 여부, 참가자 수에 대한 인지 여부 등으로 구성되며 이는 입찰자의 전략 형성에 중요한 조건이 된다. 경매 형식은 미술품처럼 점차 가격을 올리면서 진행되는 영국식 경매, 생물처럼 가격을 내리는 네덜란드식 경매, 부동산 매각처럼 입찰가를 비공개로 제출하고 최고가가 낙찰받는 일등가격 밀봉입찰, 건축사업처럼 입찰가를 몰래 써내되 최고가 제안자가 두 번째 높은 가격으로 낙찰받는 이등가격 밀봉입찰 방식으로 나뉜다. 전략적 행동은 균형전략과 수익 등가 정리를 포함하며, 경매 방식 간 기대 수익과 효율성 비교를 통해 경매 설계에 유용한 기준을 제공한다. 최저입찰가격은 판매자가 수익을 극대화하고 비효율적인 거래를 방지하기 위해 설정하는 전략적 기준가격으로, 공공자산 거래에서 핵심적인 설계 요소로 간주된다.

공동입찰여부, 경쟁수준임을 확인하였다. Hüttel et al.(2013)은 2003~2010년간 독일의 농지 경매 700건을 분석해 토지면적, 토질, 경작비, 입찰수, 현지구매비율, 지역경제, 가축밀도, 거주밀도, 평균 토지가격, 지역(더미변수) 등이 낙찰요인임을 밝혔다. Chow et al.(2014)은 1993~1997년간 싱가포르의 경매 145건을 분석하여 입찰방식의 효과가 낙찰요인임을 확인하였다. Agarwal et al.(2018)은 1990~2014년간 248건의 경매를 분석하여 녹지면적, 정년여부, 도심거리, 지하철 근접도, 입찰횟수, 공동입찰여부 등이 낙찰요인임을 확인하였다. 정승영 · 최인호(2019)는 1912~1916년간 한국의 농지 경매 135,923건을 분석하여 유찰횟수, 위치, 시기, 금리 등이 낙찰요인임을 확인하였다. 문혜정 · 조남욱(2024)은 2017~2023년간 임야 공매 8천여건을 분석하여 머신러닝 알고리즘별로 낙찰가격을 예측하여 MAPE(mean absolute percentage error)를 최대 4.26%까지 달성하였다.

건물 관련 낙찰가격 연구는 Ong(2006)이 한국의 주택 경매 1,281건을 분석하여 낙찰요인이 경매참여도와 부동산 유형임을 확인하였다. 임의택 · 이호병(2017)은 2014~2016년간 수도권 아파트 경매 196건을 분석하여 면적, 입지, 유찰횟수, 응찰자수, 채권유형 등 낙찰요인을 밝혔다. 김경태 외(2019)는 2006~2012년간 한국에서 자기주택을 낙찰받은 84건을 분석하여 자기자본비율, 총자산규모, 부동산업종, 대출비중, 대출금수익률 등 낙찰요인을 확인하였다. 김선아 · 전해정(2020)는 2002~2019년간 서울지역의 아파트 경매를 분석하여 매매가격지수, 낙찰률, 회사채 수익률, 소비자물가지수, 주거용건축허가현황 등이 낙찰요인을 확인하였다. Kang et al.(2020)은 2013~2017년간 서울 아파트 경매 9,435건을 분석하여 입찰자수, 경매소요기간, 면적, 미납관리비, 임차정보, 지상권, 교통, 물가, 금리 등 낙찰요인을 확인했다. 김도균 · 정재호(2021)는 2006~2020년간 서울의 아파트의 매매가격과 경 · 공매 거래정보간 상호관계를 밝혔다. Rhee et al.(2021)는 2010~ 2020년간 한국의 아파트 경매 111,232건을 분석하여 위치, 브랜드, 방수, 총가구수, 층, 법적권리, 경제지표 등 낙찰요인을 밝혔다. 류슬기 외(2021)는 토지면적, 건물면적, 강남구여부, 지하철과의 거리가 2020년까지 한국의 종전부동산 공매의 낙찰요인을 밝혔다. 홍일석 · 박문수(2021)는 2010~2020년간 물류창고 경매 552건을 분석하여 도심간 거리, 접면도로폭, 층고, 면적, 건폐율, 유치권, 경매방식 등 낙찰요인을 확인했다. 전해정(2023)은 2009~2023년간 매각된 서울 지역 상가 경매의 낙찰가 대상 시계열 분석을 수행하였다. 이진우 · 오세준(2023)은 2017~2021년간 서울의 아파트 경매 1,721건을 분석 · 예측하였다.

경 · 공매 부동산이 아닌 일반 부동산을 대상으로 한 자연어처리 기반의 부동산 가격 연구를 수행순서로 보면 다음과 같다. Sun et al.(2014)은 2012~2014년까지 신나통신의 부동산 기사를 분석하여 베이징, 상하이, 청두, 항저우 지역의 부동산 가격의 변동지표와의 상관성을 분석하였다. 박재수 · 이재수(2019)는 텍스트 마이닝으로 2010~2017년간 신문기사를 분석하여 도심권을 제외한 서울권역의 소형아파트에만 긍정적 뉴스가 영향

 https://doi.org/10.30902/jrea.2025.11.1.109

을 미치는 것을 확인하였다. Guo et al.(2020)은 2011~2017년 기간 동안 29개 주요 키워드에 대한 바이두 통계지수를 월별로 조회하여 상하이의 중고 주택 가격 예측에 활용하였다. Zhou et al.(2019)는 미국의 임대주택광고사이트 Craigslist에서 2018년 4~12월까지 애틀란타의 임대 부동산 게시물을 분석하여 주택 임대가격을 예측하여 MAE(mean absolute error) 145.4를 달성했다. 이재수 · 박재수(2020)는 2012~2018년까지 지상파 3사의 부동산 관련 기사 중 무작위 추출한 9,600개 내용과 KB국민은행의 서울시 아파트 매매가격지수와의 상관성을 분석하였다. Rajeshwaran et al.(2021)은 인도 부동산 시장에서 머신러닝 기반 광고 분류 기법을 활용하여 PropTech을 적용한 주택 가격 예측 방법을 연구했다. Zhu et al.(2023)은 2012년부터 2018년까지 Weibo에 게시된 140자 이하의 단건문자 88만건을 분석하여 베이징, 상하이, 선전, 광저우 지역의 주택시장 심리지수를 구현하였다. Bushuyev et al.(2024)는 미국 휴스톤의 대표부동산 웹 사이트 Redfin의 9,260개의 게시글이 아파트의 임대가격에 미치는 영향을 분석하여 예측오차 MSE를 13.4% 감소시켰다. 김수아 외(2024)는 뉴스 텍스트에 대한 자연어처리를 통해 감성 분석을 수행하고 부동산 가격을 예측하였다. 이연동 외(2024)는 BERTopic을 활용해 부동산 관련 언론기사의 주요 토픽을 추출하여 아파트 실거래가격지수와 비교해 상관관계를 분석하였다.

3. 본 연구의 차별성

본 연구는 2017년 이후 전국 공매물건 중 63%를 차지하는 임야, 전, 답, 대지 관련 자료를 수집하여 충분한 표본의 대표성을 확보하였다. 시간 흐름에 따른 영향을 분석하기 위해 8년간 종단적 자료를 활용하였으며, 공간적 특성을 반영하기 위해 시군구별 국내외 남녀인구, 주택유형별 가구 수, 세목별 지방세 등 사회경제적 지표들을 독립변수로 구성하였다.

기존 자연어 처리 연구들은 뉴스, 방송, SNS 등 부동산 가치에 대한 간접적 텍스트만을 분석하였으나, 본 연구는 토지가격 산정의 직접적 근거가 되고, 입찰참여자의 의사결정에 주요 참고 정보가 되는 감정평가서의 문장을 분석 대상으로 선정하였다. 이를 통해 부동산 가치의 정량적 요인과 정성적 요인을 통합해 분석하였다.

충분한 표본 규모와 포괄적인 변수 구성, 그리고 실제 감정평가 문장을 활용한 연구의 결과는 공매물건의 낙찰가격 예측에 국한되지 않고, 일반 부동산의 가치평가에도 적용 가능한 높은 외적 타당성을 제공할 것으로 기대된다.

Ⅲ. 연구 설계

1. 주요 변수 및 자료의 수집

분석 대상 자료는 임야, 전, 답, 대지 지목 토지 대상의 공매정보와 관련 경제지표이다. 공매정보는 공공데이터포털에서 제공하고 있는 온비드의

OpenAPI를 통해 수집하였다.[2] 공매 관련 경제지표는 통계청에서 수집하였다. 수집 대상별 수집건수, 수집자료, 추출변수를 정리하면 〈표 1〉과 같다. 전통적인 토지가격 연구에서는 시점의 변화가 중요한 설명 요인으로 간주된다(Alston, 1986; Titman, 1985). 그러나 기존의 토지 경·공매 가격 예측 연구에서는 매각 시점이 독립변수로 충분히 반영되지 않았으며, 이는 자료 수집의 한계에서 비롯된 것으로 판단된다. 본 연구는 2017년부터 2024년까지 장기간의 공매 사례를 대상으로 하며, 특히 그 기간 내에 팬데믹과 같은 비정상적 외부 충격이 존재하였기 때문에, 입찰년월을 시간성 독립변수로 포함하여 분석을 수행한다.

기존의 토지 경매 가격 예측 연구에서는 도심간의 거리가 토지가격에 영향을 미친다는 결과를 제시하고 있다(홍일석·박문수, 2021; Agarwal et al., 2018; Ooi et al., 2006). 그러나 국토 면적이 좁은 한국의 경우, 도심과의 거리보다는 행정구역 단위인 시·도 및 시·군·구를 통해 중심지 접근성을 판단하는 것이 보다 현실적이다. 또한, 『지방자치법』 제10조에 따르면 행정구역의 구분 기준은 인구를 기반으로 하므로, 인구는 중심성의 간접 지표로 활용될 수 있다(행정안전부, 2024). 이에 따라 본 연구는 기존 토지가격 연구(서교, 2005; 이창로·박기호, 2013; Orford, 2000)를 참고하여, 해당 지역의 인구와 가구 수를 독립변수로 포함한다.

감정평가 실무기준(국토교통부고시 제2023-522호 제3.3항, 제5.6항)에 따르면, 국·공유재산 및 공시지가 대상의 감정평가는 위치, 형상, 환경 등 토지의 객관적 가치 형성에 영향을 미치는 개별적인 요인을 고려하여야 한다. 그러나 기존 연구에서는 이러한 요인이 감정평가서 내 문장 형태로 기술되어 있음에도 불구하고, 이를 독립변수로 분석에 활용하지 않았다. 본 연구에서는 이 한계를 보완하고자, 감정평가서에 텍스트 마이닝을 적용하여, 각 매각 물건별 해당 내용의 포함 여부를 변수화하고 독립변수로 적용한다.

2. 분석 방법 및 전처리

정량변수와 정성변수의 분석절차는 〈표 2〉와 같다. 정량변수의 전처리 및 분석방법은 공매 관련 정보의 중복을 제거하고 물건별 연관성을 확인하기 위해 정규화를 적용하여 테이블로 분리한 후 RDB로 저장한다. 저장은 MySQL에 하는데, 이때 물건별[3] 변수, 물건별 집계[4] 변수, 경제환경[5]

2) 공매의 토지, 권리, 매각 관련 자료는 공공데이터포털에서 OpenAPI 형태로 제공되는 웹화면에서 해당 데이터를 수집하였다. 경제 환경 관련 자료는 통계청에서 문자형태로 제공되는 파일(*.csv)을 다운로드 받았다. 감정평가서 정보는 공공데이터 포털에서 OpenAPI 형태로 제공되는 웹화면에 나타는 파일의 URL 경로를 통해 다운로드 받았다.

3) 입찰년월, 토지지목, 면적, 감정가격, 지분소유여부.

4) 임금체불유무(채권 중 임금체불 관련 권리가 있는지 여부), 근로복지공단에서 등기한 채권의 유무 여부, 권리기관유형수(카드사, 은행사, 보험사, 근로복지공단, 세무서 등 채권을 주장하는 기관의 유형 수), 권리기관기관수, 배분요청채권수(공매 진행 과정에서 해당 부동산에 대해 채권자들이 청구한 채권의 수), 배분요청채권금액, 공매소요일수(공매의 매각 시작일부터 낙찰결정까지 소요된 기간), 입찰·유찰 건수, 매각정보의 조회·취소·미납 건수, 지역낙찰(가격)비율(낙찰가격/감정가격)·건수, 최저입찰가.

5) 전월·전분기·전년도의 지가변동비율, 전년도 시군구별 내·외·국인 남·녀 인구, 전년도의 시군구별 주택유형(단독주택, 아파트, 연립주택, 다세대주택, 비거주용건물내주택, 주택이외의거처)별 가구수, 전년도의 시군구별 세목별(취득세, 주민세, 지방소득세, 재산세) 지방세.

〈표 1〉 독립변수의 선정근거 및 자료의 수집

구분	독립 변수	선정 근거	자료의 출처 / 건수
토지	면적	류슬기 외(2021); 문혜정 · 조남욱(2024); 홍일석 · 박문수(2021); Agarwal et al.(2018); Hüttel et al.(2013); Kang et al.(2020); Ooi et al.(2006)	물건목록: 26,891 https://bit.ly/40VbDy0 기본정보: 26,891 https://bit.ly/4aEXWGw
	감정가격	김도균 · 정재호(2021); 류슬기 외(2021); 문혜정 · 조남욱(2024); Ong(2006)	
	지분소유여부	문혜정 · 조남욱(2024)	
권리	임금체불유무, 근로복지공단 채권여부, 권리기관유형수, 권리기관기관수, 배분요청채권수, 배분요청채권금액	문혜정 · 조남욱(2024)	권리정보: 69,596 https://bit.ly/40InjCY 배분요구: 222,167 https://bit.ly/40EAYuz 임차정보: 1,190 https://bit.ly/3EkrQE6
	배분요청합계	김경태 외(2019); 김선아 · 전해정(2020); 문혜정 · 조남욱(2024); Rhee et al.(2021)	
	임차유무	문혜정 · 조남욱(2024); Kang et al.(2020)	
매각	공매소요일수, 입찰 · 유찰 건수	문혜정 · 조남욱(2024); 정승영 · 최인호(2019); Agarwal et al.(2018); Kang et al.(2020)	입찰이력: 263,736 https://bit.ly/4jlOjuz
	공매 조회 · 취소 · 미납 건수, 지역낙찰비율 · 건수, 최저입찰가	문혜정 · 조남욱(2024)	
	입찰년도, 입찰년월	신규(Alston, 1986; Titman, 1985)	
경제 환경	취득세, 주민세, 지방소득세, 재산세, 총 인구, 가구수	문혜정 · 조남욱(2024); Hüttel et al.(2013)	통계청: https://kosis.kr/ 지방세: 2,080 인구수/가구수: 2,784 지가변동율: 28,235
	시군구별 내 · 외국인 남 · 여 인구, 시군구별 종류별[6] 가구수	신규 서교(2005); 이창로 · 박기호(2013); Orford (2000)	
	전월 · 전분기 · 전년 지가변동율	문혜정 · 조남욱(2024); Hüttel et al.(2013); Ooi et al.(2006)	
감정 평가서	맹지여부, 분묘여부	문혜정 · 조남욱(2024)	감정평가서: 26,365 https://bit.ly/4hvJxyO
	위치 · 환경, 교통, 형태 · 지세, 이용상황, 도로상황, 토지이용계획, 제시외 물건, 공부와 차이, 기타 등에서 추출한 권리, 감정, 가격, 교통, 주거, 농업, 묘지 관련 변수	신규(국토교통부, 2023)	

변수를 계산하여 물건번호를 기본키로 한 하나의 테이블에 저장한다. 지역별 변수는 모두 매각물건의 소재지의 시군구와 동일한 지역 기준이다.

정성변수는 감정평가서를 이미지로 변환 후 OCR로 문자 추출과 물건번호별 문장을 정렬로 추출하고, 다시 EXCEL 파일로 저장 후 khcoder를 이용해 텍스트마이닝으로 주요 맥락을 정량변수로 추출한다. 감정평가서의 내용을 정량화하는 방법은 Ⅳ장에서 자세히 설명하겠다.

변수통합 단계에서는 정량 변수와 정성변수를

6) 단독주택, 아파트, 연립주택, 다세대주택, 비거주용건물내주택, 주택 이외의 거처.

〈표 2〉 변수유형별 분석절차

유형 절차	정량변수 연속형 수치변수	정성변수 더미형 이항변수	
전처리	• 공매물건 자료 정규화 • 경제환경 자료 정규화 • RDB로저장 ※ 도구: MySQL	• 감정평가서 이미지 변환 • 감정평가서의 문장 추출 • 물건번호별 문장 저장 ※ 도구: OCR, Excel	
변수추출	• 물건·집계별 변수 계산 • 경제환경별 변수 계산 ※ 도구: MySQL	• 토픽분석 및 용어사전정의 • 군집 기반 정성 변수 추출 ※ 도구: Khcoder	
변수통합	• 물건별 수치변수와 정성변수를 통합하여 저장 • 물건번호별 분석용 테이블 구성 ※ 도구: MySQL		
전처리	• 분포검정: 변수들의 정규성 검증 및 log 치환 • 변수표준화: 독립변수의 표준화계수 도출 • 주성분분석: 주요 변수 도출 및 차원 축소		※ 도구: RStudio
분석	• IV, 2SLS: 낙찰가격의 인과 관계 분석		
진단	• 모형진단: 다중공선성, 자기상관성, 이분산성, 이상치, 잔차분석		
평가	• 성능평가: adj.R^2, RMSE, MAE, MAPE		

주 : IV, instrumental variable; 2SLS, two-stage least squares; RMSE, root mean square error; MAE, mean absolute error; MAPE, mean absolute percentage error.

물건번호별로 결합(join)하여 분석용 테이블을 구성하고, 인과관계분석을 위한 하나의 파일(.csv)로 저장한다.

전처리 단계에서는 먼저 변수들의 정규성 검증을 수행하여 극단적으로 치우치거나 분포가 넓은 변수들을 log로 치환한다.[7] 다음 독립변수를 표준화하여 표준화 회귀계수를 도출하고, 이를 통해 변수 간 영향력을 비교하였다. 또한, 다양한 종류의 독립변수 중 주요한 변수를 도출하고 분석 차원을 축소하기 위해 주성분분석(principal component analysis, PCA)을 수행하였다(Kumar et al., 2018; Mostofi et al., 2022).

분석 단계에서는 공매의 낙찰가격을 예측하기 위한 회귀분석을 수행하였으며, 이는 입찰회차별 및 지목별로 수치형 변수만을 활용한 정량모형과 감정평가서 기반 정성변수를 통합한 통합모형으로 구분하여 분석하였다. 특히 독립변수 중 최저입찰가는 감정평가액을 기준으로 설정되며, 감정평가액은 궁극적으로 시장가치를 추정하려는 시도이므로, 최저입찰가와 낙찰가격 간에는 내생성 문제가 존재할 수 있다. 이에 따라 본 연구는 도구변수법(instrumental variable, IV)과 2단계 최소제곱법(two-stage least squares, 2SLS)[8]을 적용하여 낙찰가격에 대한 인과관계를 통계적으로 검증하였다.

분석모형의 진단 절차는 다음과 같다. 먼저, 다중공선성은 독립변수 간의 높은 상관관계를 점검[9]하여 회귀계수 추정의 불안정성을 방지한다. 자기상관성 진단은 회귀모형의 오차항들이 시간이나 순서에 따라 서로 상관되지 않고 독립적인지

7) 종속변수의 정규성 검정 결과, 낙찰가격의 원화(KRW) 단위 값은 정규분포를 따르지 않았으나, 자연로그로 변환한 결과 정규성이 확보되었다. 이와 같은 이치로 낙찰가격, 감정가격, 최저낙찰금액, 배분요청금액 등 모든 가격 관련 변수들에 로그변환을 실시하여 변수 간 범위를 조정하였다. 독립변수 중 재산세, 주민세, 취득세, 지방소득세 등 지방세도 단위가 원(KRW)이나 이미 천 원단위로 절사되어 있고, 회귀분석 적용 시에도 원단위로 학습했을 때 결정계수와 예측오차가 적기 때문에 지방세 관련 변수는 log를 취하지 않고 사용함.

8) IV는 내생성이 의심되는 변수의 편향을 제거하기 위해 외생적 도구변수를 활용하는 추정 방법이며, 2SLS(two-stage least squares)는 이를 구현하는 방식으로, 먼저 내생변수를 도구변수로 예측 값을 이용해 최종 회귀를 수행.

9) VIF(variance inflation factor) 함수를 회귀모형에 적용해 VIF 계수가 2 이상인 독립변수 제거.

 https://doi.org/10.30902/jrea.2025.11.1.109

를 확인[10]한다. 이분산성 여부는 잔차의 분산이 일정한지를 확인[11]하는 방식으로 검토하며, 이상치는 잔차 및 영향력 지표를 통해 탐지[12]한다. 마지막으로, 잔차 분석을 통해 정규성, 분산 안정성, 자기상관 등을 종합적으로 점검[13]한다.

성능평가는 학습한 모델을 기초로 2024년 매각된 공매를 대상으로 측정한다.[14] n개의 낙찰가격의 실제값(y)과 예측값($\hat{y}$)의 차이에 대한 예측성능을 평가하기 위한 지표는 수정된 결정계수(adj. R^2),[15] 평균제곱근오차(root mean square error, RMSE),[16] 평균절대오차(MAE),[17] 평균절대비율오차(MAPE),[18] 낙찰성공건수,[19] 낙찰성공확률[20]이다.

3. 분석대상과 실험계획

수집, 전처리에 성공한 자료 24,047건 중 2017년부터 2023년까지 매각된 압류재산[21] 토지 21,515건의 자료를 학습자료로 사용하고, 2,532건은 2024년에 매각된 자료로써 시험대상이다. 공매는 매각 초기에 입찰자가 없거나 낙찰이 되지 않는 경우 유찰되고 다음 입찰회차의 최저입찰가격은 10%씩 낮아진다. 유찰은 기존 입찰가격에서 매수자가 없거나, 해당 가격이 시장에서 수요를 반영하지 못하고 있거나, 유치권이나 임차인 등 복잡한 권리관계 등이 있는 경우가 많다. 따라서 입찰 초기에 감정가 수준으로 낙찰되는 물건과 유찰이 많이 진행된 후 감정가보다 낮은 가격으로 매각되는 물건은 낙찰요인이 매우 상이하다. 따라서 입찰이 진행되는 상황에 따른 낙찰의 주요 요인을 선별하기 위해 전체 자료와 입찰 1회, 2~6회, 7회 이후 자료를 2024년 전후로 분리하여 분석한다(문혜정 · 조남욱, 2024).

10) 공매년월을 기준으로 시계열화 한 잔차를 대상으로 ACF/PACF 분석 및 ARIMA 모형을 적용 후 Durbin-Watson 검정.

11) bptest(studentized Breusch-Pagan test) 함수를 회귀모형에 적용하여 p-value가 0.05보다 작으면 이분산성을 의심하며, 이 경우 coeftest 함수를 사용하여 통계적으로 유의하지 않은(p-value〉0.05) 독립변수를 제거.

12) influencePlot 함수를 회귀모형에 적용하여 이상치를 탐지하였으며, 감정가격 확정 이후 개발 등 외부 요인의 변화로 인해 발생한 공매 낙찰가격의 극단적인 사례도 분석 대상의 현실적 반영을 위해 제거하지 않고 모형에 포함.

13) 회귀모형의 잔차에 ks.test 함수를 적용해 p-value가 0.05 미만일 경우, 잔차가 정규성을 충족하지 않는 것으로 판단하며, 이 경우 변수 정규화(log 변환 등), 이상치 제거, 강건 회귀(robust regression) 등의 방법을 통해 분석을 재수행.

14) 성능지표 RMSE, MAE, MAPE 계산 시엔 쉽게 이해하기 위해 가격변수에 지수를 취해 원(KRW) 단위로 계산.

15) 수정된 결정계수(adj. R^2=1-((1-R^2)*(n-1)/(n-k-1), R^2=((1 − ($\sum(y-\hat{y})^2$)/($\sum(y-\bar{y})^2$))))모델에 포함된 변수 수를 고려하여 과적합을 방지하고 설명력을 보정한 결정계수.

16) RMSE($\sqrt{\sum(y-\hat{y})^2/n}$): 예측값과 실제값 사이의 오차 제곱 평균의 제곱근으로, 예측 오차의 평균적인 크기를 나타내는 지표.

17) MAE($\sum | y-\hat{y} |)/n$) 예측값과 절댓값 차이의 평균, 특잇값이 많으면 유용, 낮을수록 좋음.

18) MAPE($(100/n)\sum_{i=1}^{n} | (y-\hat{y})/y |$)는 평균적인 절대 오차의 비율로써, 예측한 낙찰가격의 범위에 의존하지 않고 모델의 성능을 비교할 수 있다. MAPE는 0에서 1사이 값이고 낮을수록 예측성능이 좋다.

19) 예측가격이 실제 낙찰가격보다 같거나 높은 경우 낙찰에 성공한 것으로 간주.

20) 전체 시험건수 중에서 낙찰에 성공한 건수의 비율(%)을 계산.

21) 공매종류 압류재산, 국유재산, 유입자산, 수탁재산 중 낙찰가격 예측이 어려운 압류재산을 대상으로 분석을 수행.

Ⅳ. 분석 내용

1. 감정평가서 분석 및 정성변수 정의

1) 감정평가서의 맥락분석

〈표 3〉은 감정평가서에서 정성변수를 추출하는 의미망분석 과정이다. 감정평가서 26,365건에서 추출한 문자를 대상으로 토큰분석을 수행한 결과, 문단 49,472건, 문장 98,680건, 단어 285,351건이 추출되었고 이 중 유의미한 단어는 1,642건이며 감정평가서에 총 17,229번 등장하였다. 맥락분석을 위해 먼저 맹지, 토지이용계획, 농지취득자격증명원 등 전문용어를 선별 후 개별 단어로 구분되는 이해관계인, 중앙선, 목장용지 등의 복합어를 다시 정의하였다. 이때 특수문자, 어근어미, 접속사, 공백 등 불용어는 제거하였다.

토큰분석에서 선별된 단어 1,642건만으로 의미망분석을 하며 1단계와 같이 군집분석을 수행하여 교통환경, 농업환경, 권리관계, 묘지환경 등 주요 분류를 선별하였다. 2단계는 감정가격 대비 낙찰가격의 비율을 10단계[22]로 구분하여 의미망분석을 수행하여 낙찰가격의 비율에 따라 분포된 단어나 문구를 시각화하였다.

3단계는 낙찰가격의 비율이 높은 8단계 이상 의미망에 등장하는 단어나 문구는 교통편의. 농업유리, 가격인상 분류의 정성변수의 구문안에 정의한다. 또한, 같은 3단계에서 낙찰가격의 비율이 낮은 3단계 이하의 의미망에 등장하는 단어나 문구는 교통불편, 농업불편, 맹지, 활용제약, 권리관계, 묘지문제 등 분류의 정성변수의 구문안에 정의한다. 맥락별 포함문구는 전체 감정평가서에서 해당 문구가 얼마나 독특하게 등장했는지 측정하는 TF-IDF[23]기준으로 선별한다.

2) 이항 정성변수 적용의 타당성

일반적인 텍스트마이닝은 정성변수에 해당하는 포함문구 대비 분석대상(해당 물건의 감정평가서)에 포함된 문구의 개수를 계산하여 연속형 수치변수로 계산하나 본 연구에서는 1과 0으로 구성된 이항변수의 형태로 정성변수를 추출한다. 감정평가서는 정성변수에 해당하는 문구 하나만 있어도 해당맥락이 유의미하다고 보기때문이다. 예를 들어 교통편의 정성변수를 보면, '대중교통(편리|양호|우수)[24]'부터 '지하철'까지 35가지의 문구 중 하나만 있어도 교통이 편리하다는 것을 의미한다. 감정평가서는 하나의 토지에 대해 여러 가지 표현으로 교통사정을 중복해서 표현하지 않기 때문이다.[25]

22) 낙찰가격에서 감정가격을 나눈 값을 백분율(%)로 환산한 값으로써 01_0~39, 02_40~49, 03_50~59, 04_60~69, 05_70~79, 06_80~89, 07_90~99, 08_100~109, 09_110~119, 10_120 이상 10단계의 백분위수로 구분.

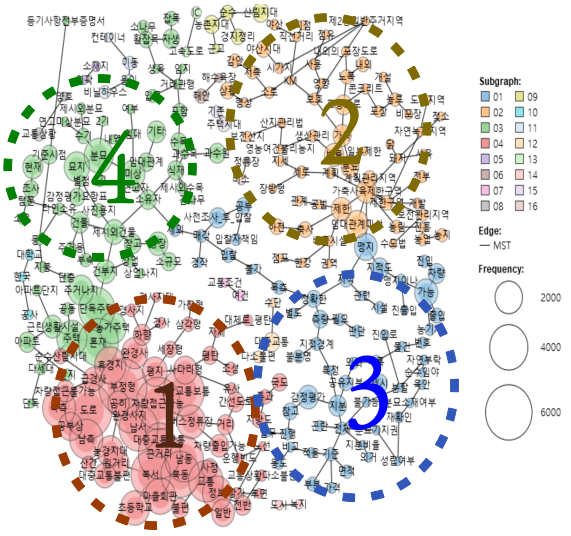
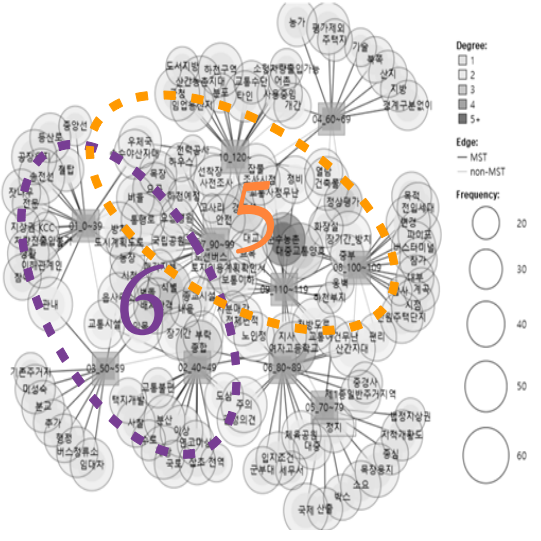
23) TF-IDF(term frequency-inverse document frequency)는 문서 내에서 단어(또는 문구)의 중요도를 평가하는 통계적 가중치로, 공매물건에 해당하는 감정평가서 내 단어의 빈도(TF)와 전체 문서 집합에서의 단어 희소성(IDF)을 결합한 지표(TF-IDF(t,d)=TF(t,d)×IDF(t)=(n(t,d)/N(d))×log(D/df(t)))이다. 여기서 n(t,d)는 문서 d에서 단어 t의 출현 빈도, N(d)는 문서 d의 전체 단어 수, D는 전체 문서 수, df(t)는 단어 t가 출현한 문서 수를 의미한다. 이를 통해 특정 문서에서 자주 등장하면서 전체 문서군에서는 희소하게 등장하는 단어에 높은 가중치가 부여된다.

24) 대중교통편리, 대중교통양호, 대중교통우수 세가지를 정규표현식으로 표현하면 대중교통(편리|양호|우수)이다.

25) 제약사항의 경우 맹지, 지상권등의 권리관계, 송전선 등 의 제약사항, 제시외 물건 등 다양한 내용을 포함하고 있기 때문에 제약조건에 따라 변수를 최대한 세분화 하여 서로 중복이 없도록 포함 문구들을 분리하였다.

 https://doi.org/10.30902/jrea.2025.11.1.109

〈표 3〉 의미망 분석을 통한 주요 정성변수 추출 사례

1단계: 감정평가서 내 단어의 군집군석을 통한 맥락 파악	3단계: 맥락별 낙찰가격비율별 포함문구 정의	
	맥락	포함문구(정규표현식)
	1. 교통 무난	대중교통(편리|양호|우수), 육로통행가능, 교통(여건|상태|사정|제반)(무난|양호|우수|편리), (내외의)?포장도로, (소형차|농기계)?(접근|통행)가능(하나)?, 차량(접근|통행)(가능|용이), 왕복(2|4)차선, 차고지, 차량접근후_도보출입가능, 내외의_거리에_버스정류장, 선착장, 지방도, 지하철
	5. 가격 인상 특성	교통사정무난, 교통수단, 여객선, 대중교통양호, 어촌, 도서지방, 버스터미널, 자연마을, 복지회관, 사용중임, 우체국, 군내, 군청, 지방도로, 센터, 도로변, 분교장, 토지이용계획확인서, 수풀, 하천구역, 군청, 소형차량출입가능
	5. 긍정 어조	진행중임, 등고평탄, 교환가치, 대규모, 환가성, 채취, 정확, 용이, 편리, 양호, 미래, 판단, 유리, 편리, 가능, 우수, 무난
	2. 농업 유리	순수농촌지대, 비닐하우스, 경지정리위주, 임업경영, 어촌, 정비, 묘목, 농기계
	2. 농업 불리	휴경지, 묵전, 농막, 묵지, 방치
	3. 맹지	맹지
	3. 권리 관계	지상권, 구분지상권, 분묘기지권, 법정지상권, 저촉여부
2단계: 감정가대비낙찰가격의 비율별 감정평가서 의미망 분석	3. 활용 제약	송전선, 선하지, 송전탑, 고압, 고압송전선, 송전선, 철탑부지, 철탑, 배출시설
	3. 제시외	제시외(비닐하우스|창고부지|건부지|물건|건물|주택|창고)
	4. 묘지 문제	외관상_접근이_곤란한_위치에_분묘가_소재할_수, 묘지, 분묘, 봉분, 분묘(소재여부|기지권), (연고미상|제시외)분묘, (1|2|3|4|5|수)기
	1. 교통 불편	(일반교통|통행사정|제반교통)불편, 대중교통(다소)?(좋지못함|불편|열세|불리|대체로불편|매우불편|약간불편), (농기계접근|직접통행|육로통행)(불가|불가능|불편), (접근|통행)(불편|불가), 차량(접근|출입|통행)?(곤란|불가|불편|어려움|곤란하나), 도로(상황|조건|여건|여건열악|비포장도로|다소불편|불편), 교통(상황|여건|조건|상태|수단|시설|제반)?(불편|불리|다소불편|매우불편)
	6. 가격 인하 특성	간선, 개발지구, 과밀억제권역, 국가산업단지!!!, 군사시설, 기독교, 기술, 스타, 자치, 재생, 재지, 전역, 전자, 중소, 지방산업단지, 재확인, 천주교, 청사, 취득, 정책, 건축신고, 주식회사
	6. 비판 어조	좋지못함, 다소열악, 미성숙, 손실보상, 저촉여부, 여건불리, 급경사, 경사지세, 불가능, 불편, 미상, 미미, 방치, 열악, 불리, 어려움, 열세, 곤란, 희박

2. 기초 통계

낙찰가격 예측에 사용된 주요 변수들의 기초통계량은 〈표 4〉와 같다. 종속변수인 낙찰가격은 최소 18만 원에서 최대 약 79억 원 사이이며, 중앙값은 약 1,258만 원, 평균은 약 3,931만 원으로 하향 평균적인 분포를 보인다. 최저입찰가격은 평균 36,111천 원으로 낙찰가격보다는 낮은 수준이며, 분포의 폭도 상대적으로 좁아, 낙찰가격이 최저입찰가를 중심으로 결정되는 경향을 보여준다. 배분요청합계는 중앙값(2억)과 평균(22억)의 차이가 매우 크며, 이는 극단적인 초고액 사례(최대 689억) 때문이다. 즉, 이 변수는 오른쪽으로 긴 꼬리를 가진 비대칭 분포(positive skewness)로써, 로그 변환이 필요하다. 토지면적은 극단적으로 비대칭적 분포(오른쪽으로 꼬리)를 보임인다. 중앙값(3,609㎡)보다 평균이 훨씬 크고, 최대값은 매우 큰 값이므로 로그 변환이 필요한 변수로 판단된다. 이러한 분포 특성은 회귀분석에 영향을 줄 수 있으므로 사전에 log 변환 등 정규화 작업이 필요하다.

유찰건수는 최대 39이나 평균이 4이고 75%의 공매가 6차 이내에 낙찰된다. 조회건수는 평균 59이고, 최대 1,418건으로 유난히 많이 조회되는 물건이 있는 것으로 확인된다. 공매 평균은 1년으로 대부분 1년 이내에 낙찰되며 최대 5년에

〈표 4〉 분야별 변수의 기초 통계량

구분	변수	Min.	1stQu.	Median	Mean	3rdQu.	Max.
물건 정보	낙찰가격(천 원)	180	5,300	12,579	39,314	32,200	7,902,000
	최저입찰가(천 원)	42	4,813	11,343	36,111	29,099	7,545,249
	배분요청합계(천 원)	0	64,387	208,907	2,222,747	818,026	689,995,603
	지분공매여부	-	-	-	0	1	1
	토지면적	-	172	532	3,609	1,651	3,446,462
매각 정보	입찰년도	2017	2018	2020	2020	2022	2024
	입찰년월	2017.01	2019.01	2020.01	2021.01	2023.01	2025.01
	유찰건수	-	-	4	4	6	39
	조회건수	-	-	-	41	59	1,418
	공매소요년수	-	-	-	0	1	5
경제 환경	단독주택수	1,411	15,051	20,602	22,306	27,133	81,102
	외국인_여자	44	514	1,319	2,896	3,249	40,478
	지방소득세(천 원)	841,024	7,930,712	21,917,870	70,352,766	75,243,362	1,067,036,144
	연간지역 낙찰비율(%)	12.98	59.74	72.92	79.36	90.63	1,515.57
	전년지가 변동율(%)	-32.23	8.06	13.74	14.67	19.89	92.67
	전월지가 변동율(%)	-63.30	5.90	12.40	14.54	20.60	237.70

 https://doi.org/10.30902/jrea.2025.11.1.109

걸쳐 매각되는 것이 있다. 주택수는 중앙값과 평균이 유사한 정규분포를 띄었으며, 외국인 여자 인구수는 평균이 중앙값의 두 배 이상으로 절반 이상의 지역의 인구는 1,319명 이하이고, 일부 대도시 지역은 최대 4만 명 이상까지도 거주하고 있어 지역 간 편차가 크다. 지가변동율은 최소 -63.3%에서 최대 237.7%까지 넓게 분포하며 전체적으로 전년 대비 지가가 상승한 지역이 많았다. 연간 지역낙찰비율은 최대 1,515.6%로 매우 높은 수준까지 나타나, 일부 지역이나 시점에서 낙찰가격의 변동 폭이 매우 클 수 있음을 시사한다.

3. 정량 분석

1) 정량분석의 모형진단

다중공선성 문제를 방지하기 위해 VIF 계수가 3 이상인 변수는 제외하였다(〈표 5〉). 낙찰가격에 대한 설명력이 높은 최저입찰가(log)는 내생성이 의심되는 변수로, 이를 보정하기 위해 도구변수(IV)와 2단계 최소자승법(2SLS)을 적용하였다.

〈표 5〉 입찰회차별(左) · 지목별(右) 정량분석의 모형진단(VIF, Durbin-Watson)

구분	변수	입찰회차별 정량분석				지목별 정량분석			
		전체	1회	2~6회	7회	임야	전	답	대지
물건	지분공매여부	1.056029	1.047455	1.084954		2.119490			1.301960
	최저입찰가(log)	1.070920	1.097366	1.179997	1.475291	2.807392	1.567379	1.061130	1.598968
매각	공매소요년수		1.046063	1.046126					
	유찰건수			1.014640	1.017907				
	입찰Y	1.025490	1.574620	1.696527	1.847642		1.688211	1.660541	
	입찰YM					1.918290			1.658507
	조회건수	1.533765	1.486214	1.653312	2.134364	2.164711	2.041931	1.548771	1.574262
경제	연간지역 낙찰비	1.025568	1.046429			1.114188	1.072538	1.012297	1.121646
	단독주택수	1.533765	1.058211		1.009949	1.063990			
	외국인_여자수								1.288012
	전년지가 변동율	1.128949	1.159107	1.162641	1.166509		1.266821		
	전월지가 변동율					1.249024		1.154712	1.105019
	지방소득세			1.107545					
모형 진단	dwtest	1.98093	2.01951	1.96127	1.95113	1.98644	1.96840	1.96586	2.00666
	p-value[26]	0.07993	0.77798	0.02601	0.03753	0.36659	0.14592	0.05273	0.59727

주 : VIF, variance inflation factor.

26) 자기상관성 진단 결과, 일부 모형(2~6회차 정량분석, 7회차 이상 정량분석, 전체입찰 통합분석, 2~6회차 통합분석)의 Durbin-Watson 검정에서 p-value가 0.05 미만으로 나타났으나, 공매자료의 비정형적 구조를 고려하여 Newey-West 표준오차 기반 회귀계수 검정을 수행하였고, 통계적 유의성이 유지된 변수만을 최종 모형에 포함함.

외생변수로는 토지면적, 유찰건수, 배분요청합계(log)를 사용하였다. 토지면적은 최저입찰가의 형성에 영향을 미치는 기초 변수이나, 낙찰가격에는 구조적으로 간접적인 영향만 미친다. 유찰건수는 최저입찰가 인하의 결정 요인이지만, 과거 시점의 정보로서 낙찰가격에는 직접적인 영향력이 크지 않다. 배분요청합계(log)는 배분 절차와 관련된 변수로 낙찰 이후의 정보를 반영하지만, 낙찰가격과의 직접적인 인과관계는 제한적이라고 판단된다. 2단계 최소자승법은 먼저 최저입찰가(log)를 종속변수로 설정하고, 오차항과 무관한 외생변수들(예: 토지면적, 유찰건수)을 독립변수로 하여 1단계 회귀식을 적합한다. 이 과정에서 도출된 최저입찰가(log)의 예측값은 오차항과 독립적이므로, 2단계에서는 이 예측값을 사용해 낙찰가격을 설명함으로써 최저입찰가(log)와 낙찰가격 간의 순수한 인과효과를 추정할 수 있다.

자기상관성 여부를 확인하기 위해 입찰회차별 및 지목별 정량분석 모형 8가지에 대해 Durbin-Watson 검정을 수행한 결과, 대부분의 모형은 DW 통계량이 2에 근접하고, p-value가 유의수준 5%를 상회하여 잔차 간 자기상관이 없는 것으로 나타났다. 다만, 2~6회차 입찰자료와 임야 대상 모형은 p-value가 0.05보다 작아 통계적으로 유의한 자기상관이 나타났으나, DW 통계량이 2에 매우 근접하고 자기상관의 정도가 크지 않다고 판단되어, 모든 모형은 실무적으로 자기상관 문제가 없는 것으로 간주할 수 있다.

2) 입찰회차별 정량분석

입찰회차 구간별로 수치형 독립변수를 표준화하여 수행한 다중회귀분석 결과는 〈표 6〉의 左에 제시하였다. 분석에 사용된 회귀계수(β)는 표준화된 값으로, 변수 간 상대적 영향력을 직접 비교할 수 있다. 그 결과, 최저입찰가(log)는 공매제도의 제도적 가격 결정 기준으로서, 모든 회귀모형에서 낙찰가격에 가장 강한 영향을 미치는 핵심 변수로 확인되었으며, 표준화 회귀계수는 1.21~1.45 범위로 나타났다. 한편, 최저입찰가와 낙찰가격 간에는 내생성 문제를 보정하기 위해 IV와 2SLS를 적용하였다.

지분공매여부는 전체 분석에서는 낙찰가격에 미약한 음(-)의 영향을 보였으나, 2~6회차 입찰에서는 양(+)의 영향으로 전환되는 특이성을 보였다. 이는 지분 물건이 입찰이 반복되며 경쟁력이 상승하는 구조를 시사하며, 낮은 최저입찰가를 통해 유리한 입지의 공유지분 물건이 오히려 조기에 낙찰될 수 있음을 보여준다. 권자·채무자 간 이해관계가 명확해지고, 적극적인 입찰 참여를 유도해 낙찰가격 상승으로 이어질 수 있음을 시사한다. 낙찰가격과 최저입찰가 간의 내생성 문제를 해결하기 위해, 토지면적을 최저입찰가를 설명하는 도구변수(IV)로 활용하였다. 이 경우 토지면적은 1단계 회귀(최저입찰가 예측)에만 사용되며, 2단계 회귀(낙찰가격 예측)에서는 계수가 산출되지 않기 때문에 낙찰가격에 대한 직접적인 영향력인 회귀계수를 확인할 수 없다.[27]

27) 또한 일반선형회귀분석(ordinary least squares, OLS)을 적용한 경우에도, 토지면적은 입찰회차 1~6회, 지목이 '답' 또는 '대지'인 경우에만 통계적으로 유의하였고, 그 외 대부분의 분석에서는 유의성을 보이지 않았다. 따라서 모든 경우에 면적 단위 가격이 낙찰가격 결정에 핵심적인 변수라고 단정하기는 어렵다.

〈표 6〉 입찰회차별(左) · 지목별(右) 정량분석 결과(표준화 회귀계수 기준)

구분	변수	입찰회차별 정량분석				지목별 정량분석			
		전체	1회	2~6회	7회	임야	전	답	대지
	(Intercept)	34.35307***	80.80026***	23.16779***	28.55001***	55.25833***	31.89459***	33.23880***	32.41607***
물건	지분공매여부	0.00557*	−0.01932*	0.00681*		−0.04317***			−0.01117*
	최저입찰가(log)	1.41399***	1.32514***	1.34113***	1.31993***	1.21493***	1.45842***	1.35432***	1.29866***
매각	공매소요년수		0.02192***	0.00699***					
	유찰건수			0.00687***	0.00162**				
	입찰Y	−0.00887***	−0.03181***	−0.00337***	−0.00603***		−0.00765***	−0.00833***	
	입찰YM					−0.01921***			−0.00790**
	조회건수	0.01221***	0.07456***	0.00484*	0.01100***	0.04287***	−0.00253	0.02017***	0.02972***
경제	연간지역낙찰비	0.03448***	0.03001***			0.02422***	0.04402***	0.01872***	0.02145***
	단독주택수	−0.01447***	−0.01879***		−0.00310*	−0.00773**			
	외국인_여자수								−0.02058***
	전년지가변동율	0.00452*	0.02126***	0.00325**	0.00599***		−0.00082		
	전월지가변동율					0.01436***		0.01239***	0.01729***
	지방소득세			−0.00528**					
분석	학습건수	21,515	6,307	9,967	5,241	8,889	5,858	4,382	2,386
	시험건수	2,532	614	833	1,085	1,113	717	431	271
성능	학습 adj.R^2	0.96917	0.93978	0.99552	0.99419	0.96833	0.96927	0.98163	0.97344
	시험 adj.R^2	0.99150	0.96584	0.99832	0.99108	0.98169	0.99090	0.99125	0.99544
	RMSE(원)	26,989,198	10,684,909	5,626,225	8,438,701	23,175,714	21,851,948	4,745,583	23,646,378
	MAE(원)	5,117,298	3,505,004	1,466,735	1,414,012	5,100,596	5,511,836	1,670,209	4,281,385
	MAPE(%)	10.2712	11.4479	4.4193	4.4051	11.2762	13.0404	5.8142	10.0407
	낙찰성공건수	1,050	329	411	567	644	279	200	194
	낙찰성공율(%)	41.4692	53.5831	49.3397	52.2581	57.8616	38.9121	46.4037	71.5867

주 : RMSE, root mean square error; MAE, mean absolute error; MAPE, mean absolute percentage error.

입찰년도 변수는 전 회차에서 일관되게 음(-)의 회귀계수를 보이며, 시점이 뒤로 갈수록 낙찰가격이 하락하는 경향을 반영한다. 이는 부동산 시장의 규제 강화, 팬데믹 여파, 지방 소멸 등 구조적 경기침체 요인이 공매시장에도 영향을 미쳤음을 의미한다. 조회건수, 유찰건수, 공매소요년수는 모두 일부 회차에서 낙찰가격에 양(+)의 영향을 미치는 것으로 나타났다. 특히 조회건수는 1회차에서 표준화 회귀계수 0.07456, 유찰건수는 2~6회차에서 0.00687로 나타나, 시장 노출기간

증가나 정보 접근성의 개선이 낙찰가격 상승에 기여할 수 있음을 보여준다.

지역의 경제지표 중 연간지역낙찰비율은 모든 회차에서 일관되게 양(+)의 관계를 나타내며, 해당 지역의 낙찰활동이 활발할수록 개별 물건의 낙찰가격도 높아지는 경향을 보였다. 반면, 단독주택 수, 지방소득세는 대부분의 회차에서 음(-)의 계수로 나타나, 상대적으로 경제 규모가 작고 주거밀도가 낮은 지역에서 공매물건의 낙찰가격이 더 높게 형성될 가능성을 시사한다.

전체 입찰에 대한 예측 성능은 MAPE 10.27%, MAE 5.12백만 원으로 나타나, 전반적으로 양호한 예측력을 확보한 것으로 판단된다. 1회차의 경우 MAPE 11.45%, MAE 3.50백만 원으로 예측 오차가 상대적으로 크게 나타났으나, 2~6회차에서는 MAPE 4.19%, MAE 1.47백만 원으로 개선되어 입찰회차가 진행될수록 예측 정확도가 향상되는 경향을 확인할 수 있다.

2024년 시험 데이터셋 기반 예측 성능 검증 결과, 학습모형의 수정된 결정계수(adj. R^2)는 전 구간에서 0.97 이상으로 매우 높은 설명력을 보였다. 일반적으로 adj. R^2가 과도하게 높을 경우 과적합 가능성이 제기되나, 본 분석에서는 7회차 이상을 제외한 대부분의 회차 구간에서 시험 데이터의 adj. R^2가 학습모형보다 더 높게 나타나, 회귀모형이 과적합되지 않았으며 일반화 성능 역시 양호한 것으로 판단된다. 이러한 결과는 낙찰가격이 제도적으로 설정된 최저입찰가격을 중심으로 근소한 차이 내에서 결정되는 경매 방식의 구조적 특성이 회귀모형에 효과적으로 반영되었음을 시사한다.

3) 지목별 정량분석

지목별 수치형 독립변수에 대한 다중회귀분석 결과는 〈표 6〉의 右와 같으며, 표준화된 회귀계수를 기준으로 낙찰가격에 대한 변수들의 상대적 영향력을 비교하였다.

임야는 전체 공매물건 중 약 25%를 차지하며, 8,889건의 학습데이터와 1,113건의 시험데이터를 기반으로 분석하였고, 낙찰성공률은 55.86%를 기록하였다. 모형의 절편은 55.26이며, 예측 성능은 MAPE 11.28%, MAE 5.10백만 원으로 나타나 상대적으로 높은 오차를 보였다. 낙찰가격에 유의한 양(+)의 영향을 미친 변수는 최저입찰가(log; β=1.21493), 조회건수(β=0.04287), 연간지역낙찰비율(β=0.02422), 전월지가변동율(β=0.01436)로 나타났으며, 이들 변수는 입찰경쟁, 정보 노출, 최근의 지가 추세가 임야 낙찰가격 형성에 영향을 미친다는 점을 시사한다. 반면, 지분공매여부(β=-0.04317), 입찰년월(β=-0.01921), 단독주택(가구수; β=-0.00773)은 음(-)의 영향을 미쳤으며, 공유지분 구조나 주거 밀도가 높은 지역일수록 낙찰가격이 낮게 형성되는 경향을 보였다.

전(논) 대상 분석은 5,858건의 학습데이터와 717건의 시험데이터를 활용하였고, 낙찰성공률은 38.91%로 전체 지목 중 가장 낮은 수준이었다. 모형 절편은 31.89이며, 예측 성능은 MAPE 13.04%, MAE 5.51백만 원으로 가장 오차가 컸다. 낙찰가격에 유의한 양(+)의 영향을 미친 변수는 최저입찰가(log; β=1.45842), 연간지역낙찰비율(β=0.04402)이었으며, 조회건수(β=-0.00253)와 전년지가변동율(β=-0.00082)은 오히려 낙찰

가격에 음(-)의 영향을 미쳤다. 조회가 많이 되고 지가가 오르는 지역의 논의 낙찰가격이 하락하는 이유를 확인해볼 필요가 있다.

답(밭) 대상 분석은 4,382건의 학습데이터와 431건의 시험데이터를 기반으로 분석되었으며, 낙찰성공률은 46.40%로 임야와 유사한 수준이었다. 예측 성능은 MAPE 5.81%, MAE 1.67백만원으로 비교적 우수한 성과를 보였고, 모형의 절편은 33.24였다. 유의한 양(+)의 영향을 미친 변수는 최저입찰가(log; β=1.35432), 조회건수(β=0.02017), 연간지역낙찰비율(β=0.01872), 전월지가변동율(β=0.01239)로 나타났으며, 낙찰가격이 지역 수요와 정보 노출도에 반응한다는 점을 보여준다. 반면, 입찰년도(β=-0.00883)는 음(-)의 영향을 미쳤다.

대지 대상 분석은 2,386건의 학습데이터와 271건의 시험데이터로 분석되었으며, 낙찰성공률은 71.59%로 가장 높은 성과를 보였다. 모형의 절편은 32.41였으며, 예측 성능은 MAPE 10.04%, MAE 4.28백만 원이다. 유의한 양(+)의 영향을 미친 변수는 최저입찰가(log; β=1.29866), 조회건수(β=0.02972), 연간지역낙찰비율(β=0.02145), 전월지가변동율(β=0.01729) 순으로, 낙찰가격이 시장 정보와 단기 지가변동에 민감하게 반응함을 확인할 수 있다. 반면, 입찰년월(β=-0.00846)과 외국인여성인구(β=-0.02058)는 낙찰가격에 음(-)의 영향을 미쳤는데, 이는 외국인 여성이 거주하는 지역은 대지 가격이 낮다는 것을 의미한다.

4. 통합 분석

1) 통합분석의 모형진단

입찰회차별 및 지목별 통합분석 모형의 다중공선성과 자기상관성을 검증한 결과, 모든 독립변수의 VIF 계수가 2 미만으로 나타나 다중공선성 문제가 없는 것으로 확인되었다(〈표 7〉). Durbin-Watson 검정 결과, 전체 입찰을 대상으로한 예측모형에서는 유의수준 5%에서 약한 자기상관이 존재한 반면, 다른 모든 모형에서는 DW 통계량이 1.95~2.00 범위에 분포하여 자기상관성이 없는 것으로 판단되었다.

2) 입찰회차별 통합분석

입찰회차의 구간별로 수치형 변수만을 활용한 다중회귀분석을 수행한 종합분석 결과를 정량분석과 비교하면 〈표 8〉의 左와 같다.

정량변수와 감정평가서에서 추출된 이항형 정성변수를 통합하여 분석한 결과, 통합모형의 절편은 정량변수만을 활용한 모형 대비 약 40% 수준으로 감소하였다. 이는 정성변수의 도입이 모형의 기준선(intercept)을 하향 조정함으로써 예측의 안정성을 제고하고, 설명력의 구조를 재편하였음을 시사한다. 해당 결과는 텍스트마이닝 기법을 통해 도출된 질적 정보가 회귀모형의 설명력을 실질적으로 보완하였다는 점에서 해석된다.

통합모형에서 물건 관련 수치형 변수의 통계적 유의성을 검토한 결과, 최저입찰가(log)와 입찰년도만이 낙찰가격에 대해 통계적으로 유의한 영향력을 유지하였으며, 기타 정량변수는 모두 유의성을 상실하였다. 이는 정성변수의 도입이 기

〈표 7〉 입찰회차별(左) · 지목별(右) 통합분석 모형진단(VIF, Durbin-Watson)

구분	변수	입찰회차별 통합분석				지목별 통합분석			
		전체	1회	2~6회	7회	임야	전	답	대지
물건	최저입찰가(log)	1.868261	1.728220	1.102790	1.138380	1.127375	1.654639	1.012962	1.004478
매각	입찰Y	1.005399	1.205801	1.018869	1.127938	1.005647	1.008932	1.007620	
맥락 정성	교통편의	1.088384				1.044805			1.004137
	교통불편	1.058189					1.041668		1.005624
	긍정어조						1.048216		
	비판어조				1.011012				
	농업불리	1.530770	1.359483				1.352969		
	농업유리					1.009815			
	인상특성	1.020465	1.030770			1.023038	1.005252	1.003433	1.003181
	인하특성	1.030345				1.023035			
	권리관계	1.049815	1.057403			1.074682	1.007712		
	맹지	1.069497	1.068870						
	묘지문제	1.057463		1.083254		1.038625	1.225786	1.006473	
	활용제약	1.040044	1.018462			1.029423			
모형 진단	dwtest	1.96461	2.00208	1.94789	1.95643	1.96760	1.97434	1.96550	1.99601
	p-value	0.00463	0.52963	0.00454	0.05638	0.21426	0.19581	0.05106	0.43623

주 : VIF, variance inflation factor.

존 수치형 변수의 설명력을 대체하거나 일부 상쇄하는 구조적 효과를 내포하고 있음을 나타낸다. 최저입찰가(log)의 회귀계수는 정량분석과 통합분석 간 유사한 수준을 유지한 반면, 입찰년도의 회귀계수는 전체 입찰 기준에서 -0.00575, 7회차 이상 구간에서 -0.00261로 나타나, 정량분석 대비 약 50% 수준으로 감소하였다. 이는 정성변수의 설명력이 시간 변수의 대리효과를 흡수하거나 중복 설명함으로써 입찰년도의 설명력을 일정 부분 약화시켰음을 의미한다.

통합모형 분석 결과, 전체적으로 감정평가서에서 추출한 정성변수가 가격을 결정하는 주요한 설명변수로 도출되었다. 대표적으로 전체입찰을 대상으로한 예측모형을 보면, 감정평가서에서 추출된 정성변수 중 교통편의(β=0.00042), 인상특성(β=0.01261)은 전체 입찰구간에서 낙찰가격에 대해 통계적으로 유의한 양의 영향을 미쳤으며, 이들 변수의 회귀계수는 최저입찰가(log)를 제외한 대부분의 수치형 변수보다 높은 수준으로 관측되었다. 반면, 교통불편(β=-0.00152), 농업불리(β=-0.00043), (가격)인하특성(β=-0.00499), (법적 · 경제적)권리관계(β=-0.00354), (토지의)활

〈표 8〉 입찰회차별(左) · 지목별(右) 통합분석 결과(표준화 회귀계수 기준)

구분	변수	입찰회차별 통합분석				지목별 통합분석			
		전체	1회	2~6회	7회	임야	전	답	대지
	(Intercept)	28.05282***	51.62938***	22.54870***	21.64764***	29.70337***	32.38576***	27.75506***	16.46449***
물건	최저입찰가(log)	1.37643***	1.33601***	1.30400***	1.32434***	1.21867***	1.47630***	1.35348***	1.29121***
매각	입찰Y	−0.00575***	−0.01737***	−0.00306***	−0.00261***	−0.00656***	−0.00790***	−0.00562***	
맥락 정성	교통편의	0.00042				0.01307***			0.00888*
	교통불편	−0.00152					−0.00419***		−0.01018
	긍정어조						−0.01194		
	비판어조				−0.00377**				
	농업불리	−0.00043	−0.00734				0.00789**		
	농업유리					0.00907*			
	인상특성	0.01261***	0.01968***			0.01195***	0.01012***	0.00543*	0.01161**
	인하특성	−0.00499**				−0.00400*			
	권리관계	−0.00354*	−0.01311**			−0.00851***	−0.00691*		
	맹지	−0.00201	−0.01058*						
	묘지문제	−0.00368*		−0.00244**		−0.00407*	0.00530*	−0.00844*	
	활용제약	−0.00488**	−0.01210*			−0.00250			
분석	학습건수	21,515	6,307	9,967	5,241	8,889	5,858	4,382	2,386
	시험건수	2,532	614	833	1,085	1,113	717	431	271
설명력	학습 adj.R^2	0.97233	0.93474	0.99541	0.99409	0.96624	0.96575	0.98136	0.97236
	시험 adj.R^2	0.99331	0.97788	0.99639	0.99167	0.98296	0.98942	0.99102	0.99700
성능	RMSE(원)	18,477,571	10,575,496	17,796,367	7,582,086	22,732,259	27,217,744	5,351,841	16,481,836
	MAE(원)	4,018,965	4,301,306	2,161,646	1,435,063	4,846,496	6,813,299	1,924,287	3,810,039
	MAPE(%)	7.8998	13.4643	4.3919	4.6168	11.7915	14.3078	5.9803	11.3187
	낙찰성공건수	1,555	465	435	672	718	306	280	215
	낙찰성공율(%)	61.41	75.73	52.22	61.94	64.51	42.68	64.97	79.34
성능 향상	RMSE 절감액(원)	8,511,627	109,413	−12,170,142	856,616	443,455	−5,365,796	−606,257	7,164,542
	MAE 절감액(원)	1,098,333	−796,302	−694,912	−21,051	254,099	−1,301,463	−254,078	471,346
	MAPE 감소비율	2.3714	−2.0163	0.0274	−0.2117	−0.5153	−1.2674	−0.1661	−1.2780
	낙찰성공 증가건수	505	136	24	105	74	27	80	21
	낙찰성공 증가율	19.9447	22.1498	2.8812	9.6774	6.6487	3.7657	18.5615	7.7491

주 : RMSE, root mean square error; MAE, mean absolute error; MAPE, mean absolute percentage error.

용제약(β=-0.00488), 맹지(β=-0.00201), 묘지문제(β=-0.00368) 등은 낙찰가격에 대해 유의한 음(-)의 영향을 미친 것으로 나타났으며, 이들 변수의 회귀계수는 -0.01 미만의 범위에서 비교적 균등한 수준을 유지하였다.

모형의 예측 성능은 정량분석과 동일한 학습 및 시험 표본을 사용하여 검증하였으며, 모든 회차별 시험 표본의 크기가 500건 이상으로 통계적 일반화 가능성과 신뢰성을 확보하였다. 7회차 이후 입찰을 제외한 모든 시험데이터의 adj.R^2는 학습데이터의 adj.R^2보다 높았다. 정성변수를 추가하여 통합분석이 정량분석보다 성능이 얼마나 향상되었는지 확인하면, 실제 낙찰가 대비 예측금액의 오차 RSME는 2~6회차 입찰과 전·답 토지의 예측모형을 제외한 나머지 다섯 개 예측모형 모두 감소했고, 낙찰성공확률은 최소 2.89%에서 최대 22.15%까지 모든 구간에서 증가하였다.

3) 지목별 통합분석

지목별 통합분석 결과는 〈표 8〉의 右와 같다. 최저입찰가(log)는 회귀분석 결과에서 회귀계수(β)가 0.99에 근접할 만큼 낙찰가격에 강한 영향을 미치는 변수로 나타났다.

정성분석의 절편값이 23.16~80.80인데 반해 비목별 절편값은 16.46~51.62로 통합분석 회귀모형이 좀 더 안정적이다. 낙찰성공률은 임야가 6.65% 증가한 64.51%, 전이 3.77% 증가한 42.68%, 답이 18.56% % 증가한 64.97%, 대지가 7.75% 증가한 79.34%로 전반적으로 예측성능이 개선되었다. 정량분석 대비 통합분석의 지목별 MAPE는 전반적으로 다소 증가했으나, RMSE는 임야가 0.44백만 원 감소한 22.73백만 원, 전이 5.36백만 원 증가해서 27.22백만 원, 답이 0.61만 원 증가해서 5.35백만 원, 대지가 7.16백만 원 감소해서 16.48백만 원으로 모든 지목에서 통하분석의 낙찰성공율은 증가했으며 임야와 대지의 예측오차도 감소했으나, 전과 답의 경우 예측오차는 증가했다.

낙찰가격을 상승시키는 요인은 지목별로 상이한데(〈표 8〉), 교통이 편리(β=0.00888~0.01307)하면 임야와 대지의 가격이 상승했고, 감정평가서에 '버스터미널', '복지회관', '우체국', '군내','군청', '지방도로', '센터', '도로변', '하천구역', '군청', '소형차량출입가능' 등의 문구가 들어간 (가격)인상특성(β=0.00543~0.01195)의 경우 모든 지목의 가격이 상승했다. 농업이 유리(β=0.01161)하면 임야가격이 상승하고, 농업이 불리(β=0.01161)하면 전가격이 상승하였다. 이는 임야는 농경을 목적으로 매수하나, 전의 경우 농경이 아니라 건축이나 기타 개발을 위해 매수하는 것으로 보인다.

지목의 특성과 용도에 따른 가치 저해 요인의 차인가 명확하다. 임야의 경우 지상권, 구분지상권, 분묘기지권, 법정지상권 등 권리관계가 주된 제약사항이다. 전의 경우 불편한교통, 권리관계가 가락하락 요인이었다. 이는 농업에 불리할수록 전 토지의 가격이 향상한 것과 같은 이유로 본다. 답의 경우 지방의 묘지가 주요 문제였고, 대지는 교통 환경이 가장 중요한 결정 요인으로 확인되었다. 교통이 불편하면 전과 대지의 가격이 하락했으며, 긍정어조가 있는 경우 오히려 대지의 가격이 하락했다. 주거에 좋은 환경이 건물을 제

 https://doi.org/10.30902/jrea.2025.11.1.109

외하고 대지만 거래하는데 가격을 낮추는 영향으로 작용한 것으로 보인다

입찰회차와 지목을 구분하지 않고 2024년에 매각된 낙찰가격을 예측한 오차비율[28]을 월별로 보면, 정량분석의 경우 낙찰가격 대비 예측값의 MAPE는 작았으나 극단적인 이상치로 인해 MAE는 오히려 크게 예측되어, 실제 입찰 시 낙찰에 실패할 가능성이 커졌다. 반대로 감정평가서의 내용을 포함해 분석한 통합분석의 MAPE는 입찰회차나 지목별로 분리해서 분석한 경우 정량분석보다 0.17%~1.28%정도 크나, 전체 자료를 분석했을 경우 2.37% 오차가 감소하였다. 특히 답 지목의 경우 낙찰성공확률은 18.56% 증가하였는데 RMSE는 0.61 감소하였다.[29] 이러한 평가를 기반으로 감정평가서의 텍스트마이닝을 통해 추출된 정성변수들을 이용해 토지 특성에 따른 가격요인을 실증적으로 규명할 수 있었다.

Ⅴ. 결론

토지공매의 낙찰가격 예측결과 정량변수와 정성변수의 특징과 차이점에 따라 활용방안에 있어서 차이가 있다. 정량분석은 동일 지역의 사례가 많고 인구, 지가, 지방세 등 경제환경 관련 정보를 알고 있을 때 예측이 잘 되었다. 지역의 경제환경 변화가 해당 물건의 낙찰가격에 영향을 미쳤기 때문에, 마치 주식 차트의 변화만 보고 투자하듯이 오차가 발생할 경우 투자손실이 큰 경우가 발생했다. 정성분석의 경우 감정평가서만 보면 낙찰가격을 예측할 수 있기에 사전에 환경이나 동일 사례에 대한 정보가 없어도 분석이 가능했다. 그러나, 대량의 감정평가서에 대한 학습이 필요하므로 정량변수 분석보다 자연어 처리를 위한 작업이 추가로 필요하다. 두 변수 모두 낙찰시 입찰회차에 해당하는 최저입찰가와 입찰년월은 주요 독립변수로 작용했다. 특히 입찰년월은 입찰 진행 여부와 상관없이 입찰일자가 2017년에서 2024년까지 최근에 가까워질수록 낙찰가격비율이 낮아지는 것을 확인할 수 있었다. 이는 전국적으로 토지의 가치가 점차 낮아진다는 것을 시사한다.

2024년 여름, ChatGPT 엔진을 기반으로 한 Estate AI 서비스가 감정평가서 요약 기능을 제공하기 시작했으나(사이냅소프트, 2024), 이는 단순한 문장 요약에 그쳐 부동산 가치의 재평가나 예측 기능은 수행하지 못했다. 반면 본 연구는 감

28) 예측오차비율은 (예측가격 - 낙찰가격) / 낙찰가격로써 양수이면 낙찰성공이고, 음수이면 낙찰실패에 해당한다.

29) 예측오차가 양수인 경우는 예측가격이 실제 낙찰가격보다 큰 것으로 예측가격대로 입찰을 하게되면 낙찰에 성공하는 경우이다. 대표적으로 정량분석에서 양수의 예측오차가 큰 물건은 전남 장흥군에 소재한 대지('2023-16252-005')와 경기 안산시에 소재한 임야('2021-09568-001')인데, 예측오차는 각각 2.231, 0.524이다. 독립변수에서 영향이 큰 변수를 찾아보면 두 물건 모두 개찰일 기준으로 1년 이내에 동일지역에서 낙찰된 물건의 감정가대비 낙찰비율이 각각 1,515.57배와 40.59로 낙찰되어 실제 낙찰비율 71.92, 22.82에 비해 너무 큰 영향이 작용한 것을 확인하였다. 한편 통합분석의 예측오차는 모두 0.171 이하로써 예측성능이 높으면서도 낙찰성공확률은 72.63%로 정량분석이 50.28%인 것에 배해 20% 이상 증가했다. 예측가격이 실제 낙찰가격보다 낮은 음수인 경우 아무리 오차가 적더라도 입찰하게 되면 무조건 낙찰에 실패한다. 정량분석과 통합분석 모두에서 가장 큰 음(-)의 예측오차를 보인 물건은 충남 천안시 소재 전 토지('2023-11319-001')와 제주 서귀포시 소재 답 토지('2022-05455-001')였다. 예측오차는 정량분석에서 각각 -0.892, -0.693을 기록했으며, 통합분석에서는 -0.890, -0.671로 나타나 두 물건 모두 통합분석이 더 정확한 예측을 보여주었다.

정평가서에 포함된 전문가의 가치판단이 담긴 공식 문서로부터 텍스트마이닝을 통해 이론적 근거를 갖춘 설명변수를 추출하고, 이를 회귀모형에 통합하여 낙찰가격 결정요인을 구조적으로 해석하는 실증적 분석을 수행함으로써, 단순 예측을 넘어 부동산 가치를 정량적으로 재평가할 수 있는 분석틀을 제시한다. 이러한 기술은 급격한 경기변동이나 시장 이상징후 발생 시 효과적인 위험관리 도구로 기능할 수 있으며, 부동산 가치평가의 새로운 패러다임으로서 산업적 활용 가능성을 갖는다.

본 연구의 한계와 향후 연구과제는 다음과 같다. 첫째, 공매입찰에서 중요하게 작용하는 유치권이나 제시외 건물 등의 정보가 분석대상에서 제외되었다는 점이다. 이러한 정보는 입찰 1주일 전에 공개되는 공매재산명세서에 기재되는데, 향후 연구에서는 이를 포함하면 보다 포괄적인 분석이 가능하리라 본다. 둘째, 2017년부터 2024년까지 CPI는 97.6에서 114.2로 상승해 화폐가치가 약 24% 하락하였으며, 동일한 토지의 명목가격은 상승하는 것이 일반적이다. 그러나 본 연구의 회귀분석에서는 모든 회차에서 입찰년월 계수가 음수로 나타나, 시간이 지날수록 낙찰가격이 하락하는 경향이 확인되었다. 이는 토지 가치 또는 수요가 감소했음을 시사하며, 향후 연구에서는 금리, 환율, 부동산 정책, 팬데믹 등 거시경제 요인과 함께 CPI를 반영한 불변가격 기반의 실질 가치 분석이 필요하다. 셋째, 텍스트마이닝으로 추출한 변수들은 감정평가 전문가가 작성한 공식 문서에서 도출된 것으로, 단순한 상관관계가 아니라 부동산 가치 평가에 대한 제도적·실무적 판단이 반영된 설명변수로 기능한다. 물론 Granger 인과성이나 경로분석을 통한 인과 방향성 검토는 향후 연구에서 보완될 수 있으며, 본 연구는 해당 변수들의 통계적 유의성과 이론적 정합성에 기반해 낙찰가격 결정 요인을 구조적으로 해석하고자 하였다. 마지막으로, 본 연구에서 활용한 정량변수와 정성변수 외에도 토양성분, 지하수위, 경사도 등의 지질학적 특성과 기후변화에 따른 자연재해 위험도, 필지형상, 진입도로 등의 물리적 특성을 정량화하여 분석에 포함시킬 필요가 있다. 이를 위해 딥러닝 기반의 자연어처리 기법을 적용하여 텍스트 분석의 정확도를 높이고, 시계열 예측모형과 기계학습 모형을 결합한 하이브리드 접근이 가능할 것이다.

ORCID

문혜정 https://orcid.org/0000-0001-8265-3256
조남욱 https://orcid.org/0000-0002-3269-1497

참고문헌

1. 국토교통부. (2023). *감정평가 실무기준 [국토교통부고시 제2023-522호, 일부개정 2023.09.13.].* https://bit.ly/4cb1bXj
2. 김경태, 조원진, 노승한. (2019). 저축은행 재무구조가 경매매각가율에 미치는 영향에 관한 연구: 영업정지 저축은행의 주거용부동산 자기낙찰을 중심으로. *주거환경, 17*(1), 45-59.
3. 김도균, 정재호. (2021). 서울시 아파트 매매시장과

경매시장 및 공매시장의 상호관계 연구. *부동산학보, 85*, 84-99.

4. 김선아, 전해정. (2020). 딥러닝을 이용한 주택 경매 시장 예측에 관한 연구. *부동산경영, 21,* 7-26.
5. 김수아, 권미주, 김현희. (2024). 생성 AI기반 뉴스 감성 분석과 부동산 가격 예측: LSTM과 VAR모델의 적용. *한국정보처리학회 논문지, 13*(5), 209-216.
6. 류슬기, 신승우, 이주응. (2021). 공공기관 이전에 따른 종전부동산의 공매에 관한 실증연구. *주택연구, 29*(3), 117-134.
7. 문혜정, 조남욱. (2024). 머신러닝 기반 한국 임야 공매의 낙찰가격 예측. *지능정보연구, 30*(2), 177-194.
8. 박재수, 이재수. (2019). 아파트 매매가격과 부동산 온라인 뉴스의 교차상관관계와 인과관계 분석: 온라인 뉴스 기사의 비정형 빅데이터를 활용한 감성분석 기법의 적용. *국토계획, 54*(1), 131-147.
9. 사이냅소프트. (2024). *부동산 감정평가 AI 사례 뜯어보기.* https://bit.ly/42RN4mN
10. 서교. (2005). 헤도닉분석기법과 공간계량경제모형을 이용한 농촌지역 지가의 영향인자 분석. *농촌계획, 11*(3), 11-17.
11. 이연동, 서경규, 조영석. (2024). BERTopic을 활용한 언론기사와 아파트 실거래가격지수의 관계분석. *한국데이터정보과학회지, 35*(6), 835-842.
12. 이재수, 박재수. (2020). 방송뉴스 감성지수와 서울시 주택매매가격의 상관 및 인과관계 분석. *주택도시금융연구, 5*(2), 53-68.
13. 이진우, 오세준. (2023). 아파트 경매 낙찰가격 결정요인에 관한 연구: 서울시 5개 권역 경매관할법원을 중심으로. *부동산분석, 9*(1), 155-171.
14. 이창로, 박기호. (2013). 지가형성요인의 다수준 종단 분석. *대한지리학회지, 48*(2), 272-287.
15. 임의택, 이호병. (2017). 수도권 아파트의 경매 낙찰가율에 미치는 영향 요인 연구. *부동산학보, 69,* 116-130.
16. 전해정. (2023). 머신러닝을 이용한 상가 경매 매각가율 추정에 관한 연구. *한국콘텐츠학회 논문지, 23*(7), 120-127.
17. 정승영, 최인호. (2019). 농지 경매의 낙찰가율에 대한 영향 요인 연구. *한국지적정보학회지, 21*(3), 63-70.
18. 홍일석, 박문수. (2021). 물류창고 경매물건의 부동산 가치 형성요인에 관한 연구: 경기지역을 중심으로. *대한부동산학회지, 39*(3), 141-165.
19. 행정안전부. (2024). *지방자치법 [법률 제19951호, 시행 2024. 5. 17.].* https://bit.ly/4bTapqW
20. Agarwal, S., Li, J., Teo, E., & Cheong, A. (2018). Strategic sequential bidding for government land auction sales: Evidence from Singapore. *The Journal of Real Estate Finance and Economics, 57*, 535-565.
21. Alston, J. M. (1986). An analysis of growth of U.S. farmland prices, 1963-82. *American Journal of Agricultural Economics, 68*(1), 1-9.
22. Bushuyev, S., Bushuiev, D., Kravtsov, D., Poletaev, N., & Malaksiano, M. (2024). Machine learning model for house price predicting based on natural language text data analysis. In *Proceedings of the 6th International Workshop on Modern Machine Learning Technologies (MoMLeT-2024).* CEUR-WS.org
23. Chow, Y. L., & Ooi, J. T. L. (2014). First-price sealed-bid tender versus English open auction: Evidence from land auctions. *Real Estate Economics, 42*(2), 253-278.
24. Guo, J., Chiang, S., Liu, M., Yang, C. C., & Guo, K. (2020). Can machine learning algorithms associated with text mining from internet data improve housing price prediction performance? *International Journal of Strategic Property Management, 24*(5), 300-312.

25. Hüttel, S., Odening, M., Kataria, K., & Balmann, A. (2013). Price formation on land market auctions in East Germany: An empirical analysis. *German Journal of Agricultural Economics, 62*(2), 99-115.
26. Kang, J., Lee, H. J., Jeong, S. H., Lee, H. S., & Oh, K. J. (2020). Developing a forecasting model for real estate auction prices using artificial intelligence. *Sustainability, 12*(7), 2899.
27. Kumar, K. S., Soundarya, K., Harshitha, R., Geetha, D. E., & Suresh, T. V. (2018). Real estate data analysis using principal component analysis and 'R'. *International Journal of Pure and Applied Mathematics, 119*(15), 1535-1541.
28. Milgrom, P. (2004). *Putting auction theory to work*. Cambridge University Press.
29. Milgrom, P. (2019). Auction market design: Recent innovations. *Annual Review of Economics, 11,* 383-405.
30. Milgrom, P. R. (1985). *Auction theory* (Cowles Foundation Discussion Paper No. 1020). Yale University.
31. Mostofi, F., Toğan, V., & Başağa, H. B. (2022). Real-estate price prediction with deep neural network and principal component analysis. *Organization, Technology and Management in Construction: An International Journal, 14*(1), 2741-2759.
32. Ong, S. E. (2006). Price discovery in real estate auctions: The story of unsuccessful attempts. *The Journal of Real Estate Research, 28*(1), 39-60.
33. Ooi, J. T. L., Sirmans, C. F., & Turnbull, G. K. (2006). Price formation under small numbers competition: Evidence from land auctions in Singapore. *Real Estate Economics, 34*(1), 51-76.
34. Orford, S. (2000). Modelling spatial structures in local housing market dynamics: A multilevel perspective. *Urban Studies, 37*(9), 1643-1671.
35. Rajeshwaran, R. (2021). Proptech for proactive pricing of houses in classified advertisements in the Indian real estate market. *International Journal of Advance and Innovative Research, 8*(1), 67-70.
36. Rhee, J. T., Ahn, W. B., & Oh, K. J. (2021). Using machine learning algorithms to forecast the optimal bidding rate in apartment auctions. *Quantitative Bio-Science, 40*(1), 31-37.
37. Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy, 82*(1), 34-55.
38. Sun, D., Du, Y., Xu, W., Zuo, M. Y., Zhang, C., & Zhou, J. (2014). Combining online news articles and web search to predict the fluctuation of real estate market in big data context. *Pacific Asia Journal of the Association for Information Systems, 6*(4), 19-37.
39. Titman, S. (1985). Urban land prices under uncertainty. *The American Economic Review, 75*(3), 505-514.
40. Tse, M. K., Pretorius, F. I. H., & Chau, K. W. (2011). Market sentiments, winner's curse and bidding outcome in land auctions. *The Journal of Real Estate Finance and Economics, 42*, 247-274.
41. Zhou, X., Tong, W., & Li, D. (2019). Modeling housing rent in the Atlanta metropolitan area using textual information and deep learning. *ISPRS International Journal of Geo-Information, 8*(8), 349.
42. Zhu, E., Wu, J., Liu, H., & Li, K. (2023). A sentiment index of the housing market in China: Text mining of narratives on social media. *The Journal of Real Estate Finance and*

Economics, 66(1), 77-118.

논문접수일: 2025년 2월 15일
심사(수정)일: 2025년 3월 29일
게재확정일: 2025년 4월 11일

국문초록

본 연구는 감정평가서의 내용을 기반으로 공매 토지의 낙찰가격을 예측하는 것을 목적으로 한다. 분석대상은 2017년부터 2024년까지 온비드(onbid)를 통해 매각된 토지 24,047건(답 4,813건, 대지 2,657건, 임야 10,002건, 전 6,575건)의 물건정보, 입찰정보, 감정평가서 정보와 해당 물건 소재 지역의 인구 · 가구(2,784건), 지방세(2,080건), 지가(28,235건) 정보이다. 분석방법으로는 감정평가서에서 정성변수를 추출하기 위한 텍스트마이닝과 낙찰가격의 인과관계 분석을 위해 도구변수법(instrumental variable)과 2단계 최소제곱법(two-stage least squares)을 사용하였다. 예측성능은 adj.R^2, RMSE(root mean square error), MAE(mean absolute error), MAPE(mean absolute percentage error), 낙찰성공건수, 낙찰성공확률로 평가하였다. 입찰회차와 지목별 낙찰가격 예측 결과, MAPE는 대부분 정량분석이 우수했으나, MAE와 RMSE는 통합분석이 우수했으며, 낙찰성공확률은 모든 경우에서 통합분석이 약 2.89%~22.15% 높게 나타났다. 인근지역 공매의 낙찰가격비율 등 정량변수는 학습 가능한 기존 공매정보와 경제환경 정보가 풍부하고 경기변화가 안정적일 때 효과적이었으며, 교통여건, 감정어조, 권리관계 등 정성변수는 기존 정보가 부족하거나 경기변화가 불안정할 때 유용했다. 본 연구는 감정평가서의 내용 분석을 통한 정성적 변수 정의 방법을 최초로 제안했을 뿐만 아니라, 정량변수와 정성변수의 특성을 규명하여 상황에 따라 적합한 낙찰가격의 예측방법을 제시했다. 또한, 높은 설명력과 낮은 예측오차를 가진 가격요인을 도출하여 공매 및 일반 토지의 가치평가에 적용할 수 있을 것으로 기대된다.

주제어 : 공매, 감정평가서, 가격 예측, 텍스트 마이닝, 프롭테크