토지 보상비 추정 모델 개발 – 건설CALS데이터와 공공데이터 중심으로

이상규[O], 김진욱[*], 서명배[*]
[O]한국건설기술연구원,
[*]한국건설기술연구원
e-mail: sanggyulee@kict.re.kr[O], jukim@kict.re.kr[*], smb@kict.re.kr[*]

Development of Land Compensation Cost Estimation Model : The Use of the Construction CALS Data and Linked Open Data

Sang-Gyu Lee[O], Jin-Wook Kim[*], Myeong-Bae Seo[*]
[O]Korea Institute of Construction Technology,
[*]Korea Institute of Construction Technology

● 요 약 ●

본 연구는 토지 보상비의 추정 모델 개발을 위해서 건설 CALS (Continuous Acquisition & Life-cycle Support) 시스템의 내부데이터와 개별공시지가 및 표준지 공시지가 등의 외부데이터, 그리고 개발된 추정 모델의 고도화를 위한 개별공시가 데이터를 기반으로 생성된 데이터를 활용하였다. 이렇게 수집된 3가지 유형의 데이터를 분석하기 위해서 기존 선형 모델 또는 의사결정나무 (Tree) 기반의 모델상 과적합 오류를 제거할 경우 매우 유용한 알고리즘으로 Decision Tree 기반의 Xgboost 알고리즘을 데이터 분석 방법론으로 토지 보상비 추정 모델 개발에 활용하였다. Xgboost 알고리즘의 고도화를 위해 하이퍼파라미터 튜닝을 적용한 결과, 실제 보상비와 개발된 보상비 추정 모델의 MAPE(Mean Absolute Percentage Error) 범위는 19.5%로 확인하였다.

키워드: 건설 CALS 시스템(Construction Continuous Acquisition & Life-cycle Support System), Xgboost 알고리즘 (Xgboost Algorithm), 하이퍼파라미터 튜닝 (Hyperparmeter Tuning), 토지 보상비 추정 모델 (Land Compensation Cost Estimation Model)

I. 서론

본 연구는 건설CALS(Continuous Acquisition & Life-cycle Support)시스템에서 축적되어 있는 내부데이터와 공공데이터 기반의 외부데이터, 그리고 모델 고도화를 위한 파생 데이터로 총 3가지 데이터 영역으로 구분되어 있다. 이러한 3가지 데이터를 기반의 토지보상비를 추정하는 모델을 개발에 관한 연구이다.

우선, 건설CALS시스템은 건설사업의 기획,설계시공,유지관리 등의 모든 과정에 대한 정보를 관리 및 지원하는 정보화 체계를 의미한다. 국토교통정보화 기본계획에 의거해 정보 자원 및 데이터 관리체계 기술을 적용한 미래 정보화 추진, 건설정보화 균형 발전 도모, ICT 융복합을 통합한 스마트 건설 실현 등의 8개 중점 과제로 구성되어 추진 중이다. 국토교통부, 5개 지방국토관리청, 18개 국토관리사무소 등에서 관리 및 활용되고 있다. [1]

특히, 건설CALS시스템은 건설사업관리, 시설물유지관리, 건설인허가, 용지보상, 건설사업정보 포털, 건설기술 정보시스템으로 구성되어 있다. 이 중 용지보상시스템은 용지보상 업무의 투명성을 확보하고 효율성을 강화하기 위해 기본조서 작성, 산정 및 감정평가, 보상협의, 수용, 재결, 공탁 등 보상 업무 전체를 디지털예산회계시스템과 연계해 처리하는 시스템으로 구성되어 있다. 토지보상과 관련된 데이터 테이블은 토지보상현황정보, 토지기본정보, 토지이동정보가 있으며, 보상비가 지급된 각 필지에 대해 평가금액, 지목, 보상액에 대한 정보가 포함되어 있다. [2]

이러한 건설CALS시스템은 2008년부터 구축되어 데이터가 축적되고 있는 상황이다. 본 연구에서 사용되는 내부데이터는 2008년부터 2018년까지 누적된 전국 건설CALS 데이터를 바탕으로 활용한다.

다음으로, 외부데이터로는 표준지공시지가와 개별공시지가를 활용하였다.

국토교통부 <부동산가격 공시에 관한 법률>에 따르면, 표준지 공시지가는 국토교통부 장관이 조사 평가하여 공시한 표준지의 단위면적 당 가격을 의미한다. [3]

특히, 표준지 공시지가는 국토교통부장관이 토지이용상황, 주변

환경, 그 밖의 자연적 및 사회적 조건이 일반적으로 유사하다고 인정되는 일단의 토지 중에서 선정한 표준지에 대하여 공시한 적정가격을 의미한다. 더불어, 표준지 공시지가는 전국에 50만 개가 존재하며 감정평가사들이 표준지 공시지가를 공시하면, 이를 바탕으로 지자체에서 나머지 개별 토지에 대한 가격을 조사 및 산정하여 개별 공시지가를 공시한다. 개별공시지가는 시장 및 군수 또는 구청장이 공시한 관할 구역 안의 개별 토지의 가격을 의미한다.

따라서, 본 연구에서는 표준지 공시지가와 개별공시지가를 토지보상금액의 주요 요인으로 판단한다. 현재 국토교통부의 공공데이터로 매년 토지 공시지가는 공개되고 있으며, 전국 표준지 공시지가는 2012년부터 제공하고 있다. 본 연구에서는 외부데이터로 2012년부터 2018년까지 전국 표준지 공시지가와 개별공시지가 데이터를 활용한다.

마지막으로 본 연구에서 사용되는 파생데이터로는 표준지 공시지가를 활용하여 비교표준지의 공시지가와 시/군구 코드, 대지 구분에 대한 변수를 활용하였다.

우선, 비교표준지 공시지가는 국토교통부 "2020년도 적용 개별공시지가 조사산정지침"에 의거하여 비교 표준지 선정 모듈을 개발하였다.

아래 Figure.1과 같이 비교 표준지 선정 모듈의 상세내역은 다음과 같다. 우선, 가장 적합한 비교표준지 공시지가를 선정하기 위해서는 대상 필지가 속한 행정동을 확인한다. 이를 기반으로 표준지들의 특성과 공시지가 가격을 적용한 후, 용도지역과 토지이용상황이 동일한 표준지를 따로 분류한다. 분류된 표준지는 7가지 토지특성(지목, 이용상황, 용도지역1, 형상, 지세명, 대지구분, 도로 교통)을 기준으로 동일한 특성에 맞추어 유사도를 점수를 측정한다. 측정된 점수가 가장 높을 경우, 해당 표준지를 비교표준지로 선정한다.

Fig. 1. Comparative Standard Land Parcels Selection Module

더불어, 19자리의 지번체계(PNU: Parcel Number Unit)를 활용하여 파생 변수로 시/군/구 및 대지 코드를 추가 생성하였다.

이렇게 구성된 3가지 영역의 데이터의 총 개수는 32,216개로 구성되어 토지 보상비 추정모델 개발에 활용한다.

II. 데이터 현황

본 연구는 토지 보상비 추정 모델 개발을 위해 앞서 설명한 데이터 영역에서 총 12개의 변수를 활용한다.

우선, 독립변수는 총 11개의 데이터로 구성한다. 시계역적인 변수인 기준년도를 기준으로 비교표준지 공시지가, 기준년도 3개년 지가변동률 평균, 지목, 지형 고저, 지형 형상, 도로 접면, 기준년도, 대지구분, 용도지역 대분류, 이용상황 대분류, 시군구 코드를 활용한다.

종속변수로는 보상평가액 데이터를 사용한다. 보상평가액은 감정평가 법인 3개의 보상 평가 금액의 산술평균값을 의미한다. 표준지 공시지가 역시 단위 면적 당 가격이라는 점, 좀 더 정확한 비교가 가능하다는 점 등의 이유로 총 보상액이 아닌 법인평가산술평가액을 활용한다.

본 연구의 모델 개발을 위해 사용되는 데이터 구성은 아래 Table 1. 과 같이 정리하였다.

Table 1. Data of Organization

Name	Content (Data Format)	Classification
P_COMPAY	Appraisal and Assessment (Numerical)	Dependent Variable
PL_PAY	Comparative Standard Land Parcels in Land Price (Numerical)	Independent Variable
RATIO	Average of Rate of Change in Price of Land in 3-Year (Numerical)	Independent Variable
JIMOK	Land Category (Categorical)	Independent Variable
GEO_HL	Terrain Elevation (Categorical)	Independent Variable
GEO_FORM	Terrain Shape (Categorical)	Independent Variable
ROAD_SIDE	Road (Categorical)	Independent Variable
CALS_YEAR	Year (Categorical)	Independent Variable
CLASS	Land (Categorical)	Independent Variable
SPFC1_MAIN	Use District (Categorical)	Independent Variable
LAND_USE_MAIN	State of Use (Categorical)	Independent Variable
CTY_CD	City District Code(Categorical)	Independent Variable

III. Xgboost 알고리즘 기반의 모델 개발

본 연구의 모델 구축에 활용된 3가지 영역의 데이터를 시간적 범위는 2012년부터 2018년까지이며, 공간적 범위는 전국 전체 필지이다. 더불어, 해당 모델을 구축하기 위해 전체 데이터 32,216개에서 7:3의 비율로 Training 데이터와 Test 데이터로 분리하여 모델을 개발한다.

1. Xgboost 알고리즘

Xgboost 알고리즘은 점진적 혹은 순차적으로 결합하여 만든 부스팅 방식의 알고리즘으로 Gradient Boost 알고리즘을 기반으로 한 기계학습 알고리즘이다.

우선, Gradient Boost는 경사하강법을 사용해 성능을 개선한 기법으로 함수의 경사를 구해 경사가 낮은 쪽으로 이동시키며 극값에 이를 때까지 오류를 최소화하는 최적의 파라미터를 찾는 방법이다.[4] 하지만, 학습 성능은 좋으나, 시간이 많이 걸리는 단점이 있다.

이에 반해 Xgboost 알고리즘은 의사 결정 나무를 구성할 때 노드 단위로 병렬 처리 기법을 사용해 비약적으로 시간 측면의 성능을 개선한 알고리즘으로 많이 활용되고 있다.[5][6]

본 연구에서는 Xgboost 알고리즘을 활용하여 토지 보상비 추정 모델을 개발하고자 한다.

2. 하이퍼파라미터 튜닝

Xgboost 알고리즘에 적용되는 파라미터의 최적화를 위해서 하이퍼파라미터(Hyperparameter) 튜닝 기법을 통해 설정값을 찾는다. 하이퍼파라미터는 기계학습 기반의 모델을 학습할 경우, 설정 범주 내에서 값을 입력하는 변수이다. [7] 따라서, 모델의 적절한 설정값을 입력해주어야 보다 정확한 학습 결과가 나타난다.

하이퍼파라미터 튜닝 방식 중 Grid search는 하이퍼파라미터를 찾기까지 많은 시간이 걸리며 사용자가 지정한 설정값 이외에 다른 설정값은 찾을 수 없는 단점이 있다. [8]

하지만, 이전의 하이퍼파라미터의 결과값을 바탕으로 새로운 하이퍼파라미터를 탐색하는 방법인 베이지안 최적화 (Bayesian Optimization) 방식을 본 연구에 적용한다.[9]

아래 Figure.2와 같이, 베이지안 최적화 방식을 통해 탐색된 파라미터 값 (Subsample (1.0), Colsample_bytree (0.4485), max_depth (10.0), min_child_weight (7.0), eta (0.3397))을 확인하였다.

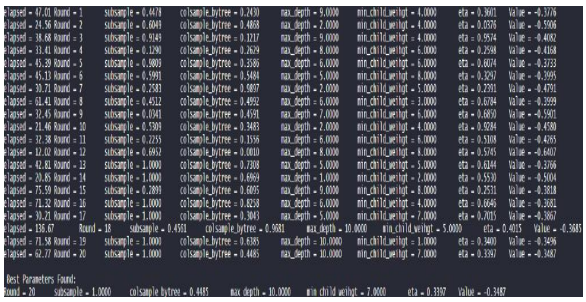

Fig. 2. Parameter Search based on Bayesian Optimization Method

IV. 분석 결과

앞서 탐색된 최적의 파라미터를 모델에 적용하여 실제 값과 예측된 값과의 오차율을 확인하기 위해 MAPE(Mean Absolute Percentage Error)을 통해 확인한다. MAPE는 크기에 의존적인 단점을 보완하기 위해 예측 값과 실제 값을 뺀 뒤 실제 값으로 나눈 비율을 의미한다.[10]

본 연구에서는 개발한 Xgboost 알고리즘 기반의 토지 보상비 추정 모델의 MAPE는 0.195608 (19.56%)이다. 다시말해, 종속변수인 실제 법인평가산술평가액의 값과 예측 값의 오차율은 19.56%로 나타났다.

ACKNOWLEDGEMENT

본 연구는 국토교통부/국토교통과학기술진흥원의 지원으로 수행되었음 (과제번호 20SCIP -C146569-03).

REFERENCES

[1] I, Chung., "Improvement Direction of Facility Management System(FMS) through Related Systems Analysis", Korea Academy Industrial Cooperation Society, Vol.16, No.10, pp.6751-6758, 2015.

[2] M. Seo., N. Kim., "Research about sharing of compensation data between owner and contractor for construction CALS improvement" The Korean Institute of Information Scientists and Engineers, Vol.37, No.1, pp.144-147, 2010.

[3] Real Estate Price Disclosure Act, Ministry of Land, Infrastructure and Transport, 2020.

[4] J. Mo., J. Yu., “Korean Web Content Extraction using Tag Rank Position and Gradient Boosting” Journal of Korean Institute of Information Scientists and Engineers, Vol.44, No.6, pp.581-586, 2017.

[5] T. Chen., C. Guestrin., “Xgboost: A scalable tree boosting system” Proceedings of the 22nd, ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.785-794. 2016.

[6] I. Kim., K. Lee., “Tree based ensemble model for developing and evaluating automated valuation models: The case of Seoul residential apartment” Journal of the Korean Data And Information Science Society, Vol.31, No.2, pp.375-389, 2020.

[7] Y. Jae., D. Ham., Y. Lee., G. Kim., "Short-term Load Forecasting Using XGBoost and the Analysis of Hyperparameters" The transactions of The Korean Institute of Electrical Engineers, Vol.68, No.9, pp.1073-1078.2019

[8] J. Jeong., J. Kim, “A Bayesian Optimization Methodology for Improving Performance of Deep Q-Networks,”in Proceedings of the KIISE Conference, pp. 773-775, 2018.

[9] J. Snoek, H. Larochelle, and R. P. Adams, “Practical bayesian optimization of machine learning algorithms,” Advances in neural information processing systems, pp.2951-2959, 2012.

[10] J. Park., S. Park., "A Study on Prediction of Attendance in Korean Baseball League Using Artificial Neural Network" KIPS(Korea Information Processing Society)

Transactions on Software and Data Engineering, Vol.6.,
No.12., pp.565-572, 2017