머신러닝 모델을 적용한 주택가격 예측 및 영향 요인 분석

백승준[O], 김준완[*], 백주련(교신저자)[*]
[O]평택대학교 데이터정보학과,
[*]평택대학교 데이터정보학과
e-mail: {nabsj[O], rlawnsdhks7[*], jrpaik[*]}@ptu.ac.kr

Prediction of Housing Price and Influencing Factor Analysis with Machine Learning Models

Seung-June Baek[O], Jun-Wan Kim[*], Juryon Paik(Corresponding Author)[*]
[O]Dept. of Data Information & Statistics, Pyeongteak University,
[*]Dept. of Data Information & Statistics, Pyeongteak University

● 요 약 ●

주택 매매에 있어서 가격에 대한 예측은 매우 중요하지만, 실거래 발생 전까지는 정확한 가격을 알 수 없다. 그렇기에 주택가격을 예측하는 많은 연구가 진행되어왔다. 주택가격을 결정하는 영향요인은 크게 주택의 내부요인과 주택의 외부 요인으로 구분되는데, 내부적인 요인 (공급면적, 전용면적, 층, 방 개수 등)에 대한 연구가 많이 진행되었다. 하지만 외부적인 요인 (위치 요인, 금융요인 등)에 대한 연구는 미비하였다. 본 연구는 주택 매수자 관점에서 가격 예측 시 외부적인 요인 역시 중요하다고 판단하여 외부요인을 적용하고자 한다. 본 논문에서 제안하는 방법은 다양한 외부요인 중 주택의 위치 정보를 활용하여, 해당 정보 기반으로 도출 가능한 데이터를 추가한다. 또한 이용량에 따른 지하철역 데이터를 추가하여 관련된 여러 영향요인들을 분석 및 적용 후 머신러닝 기반 예측 모델을 생성한다. 생성된 모델들에 주택매매 실거래 데이터를 적용하여 예측 정확도를 비교 후 높은 정확성을 보이는 모델 결과에 주요하게 영향을 끼치는 요인에 관하여 기술한다.

키워드: 주택가격(housing price), 생활사회간접자본(life social overhead capital), 지도학습(supervised learning)

I. Introduction

한국부동산원(KREB) 자료 '주간 아파트 동향'에 따르면 2021년 8월경 수도권 전체 주택가격 상승률은 0.4%를 기록했다. 하지만 2022년 12월경 수도권 전체 집값 하락률은 0.64%를 기록한다. 약 16개월 만에 상승세였던 주택가격이 급격한 내림세에 접어들고 있다. 이러한 불안정성은 주택시장에 지속적인 문제를 야기하기 때문에 가격을 예측하여 불안정성을 감소시킬 필요성이 꾸준히 대두된다. 객관적인 지표 즉 요인들에 기반을 둔 주택가격 예측을 위한 다양한 연구들이 진행되었고 여전히 진행 중이다. 주택가격을 결정하는 요인은 크게 주택의 내부요인과 주택의 외부요인으로 구분한다. 내부적인 요인에는 공급면적, 전용면적, 층, 방 개수 등이 해당하며 외부적인 요인에는 위치나 금융 관련 등이 해당한다. 선행연구들의 상당수는 내부요인을 사용한 연구들이었으며 외부요인에 대한 연구는 상대적으로 미비하다. 본 연구는 주택 매수자 관점에서 가격 예측 시 외부적인 요인 역시 중요하다고 판단하여 외부요인을 주요하게 반영하고자 한다. 여러 외부요인 중 주택의 위치 정보를 활용하여 주택가격에 영향을 미칠 만한 특성들을 추출하여 주택가격 예측 및 높은 영향요인을 분석한다.

II. The Proposed Scheme

1. Plan

주택가격 예측 및 영향요인을 분석하기 위해 역세권 데이터를 사용하였다. 역세권 데이터는 지하철역 이용량뿐만 아니라 위치정보 역시 파악할 수 있다는 장점을 갖는다. 한국부동산원(KREB)에서 제공한 역세권(500m) 이내의 실거래 데이터[1]로 2021년 10월 ~ 2022년 9월의 거래 데이터이다. 2021년 10월 ~ 2022년 8월까지의 50,395건 데이터를 학습용 데이터로, 2022년 9월 2,835건의 데이터를 테스트용 데이터로 설정한다. 표적변수는 매매가로 정한다.

외부요인 위치정보를 통해 도출할 수 있는 데이터로 주택이 위치한

지역의 평균 소득분위 특성을 추가하고, 이용량으로 가중치를 부여한 지하철역 특성, 주변 초등학교 개수, 중학교 개수 등 생활사회기반시설(life-SOC) 관련 특성을 추가한다. 참고문헌[2]에 따르면, 기존 연구는 주택가격 영향요인 중 생활 사회기반시설과의 거리만을 활용하였다. 그러나 지하철역이라도 '인천공항2터미널역'과 '강남역'의 유동 인구에 있어서 지하철역 간의 큰 차이가 발생한다. 즉, 거리뿐만이 아니라 유동인구에 따른 이용량을 고려할 필요가 있는 것이다. 유의할 점은 데이터 수집 기간 중 거래되지 않은 역세권 부동산이 있을 수 있기 때문에, 실거래 데이터에 지하철역을 바로 반영하는 방법이 아닌, 수도권 전체 지하철역별[3, 4] 승하차 유동 인구 비율을 계산[5]후 실제 거래에 가중치를 반영한다. 본 연구는 지하철역의 이용량 가중치를 반영했다는 점에서 선행연구와 차별성을 갖는다. 또한 수도권 전체를 지역으로 설정했다는 것과 주택이 위치한 지역의 소득분위를 산출하여 외부요인 반영에 있어 기존 연구들과 차별성을 두었다.

동시에 또 다른 외부요인으로 간주되는 금융관련 요인으로 계약일을 반영한 한국은행 기준금리 특성도 추가한다. 내부적인 요인인 건물의 나이, 국내 도급순위 TOP 20 건설사가 시공한 주택 유무 특성을 추가해 모델링을 진행한다. 전처리한 모델링 입력변수는 Table.1에 보인다.

2. Modeling

전처리 과정에서 표적변수의 정규성을 만족하지 않아 로그 변환을 진행하였기에 모델링 평가지표로 RMSE (Root Mean Squarrd Error) 대신 RMSLE (Root Mean Squared Log Error)로 계산한다.

Table 1. Input features

변수명	설명
지하철역	이용객 수에 가중치를 부여한 지하철역 순위
NEAR_SUBW_DIST	지하철역과 해당 주택과의 거리
소득분위	지역별 평균 소득분위로 가중치를 부여한 지역별 순위
base_rate	계약일에 따른 한국은행 기준금리
ele_school	해당 주택 500M 이내의 초등학교 개수
mid_school	해당 주택 500M 이내의 중학교 개수
park	해당 주택 500M 이내의 공원 개수
lib	해당 주택 500M 이내의 도서관 개수
FLR	주택이 위치한 층
XUAR	주택의 면적
COMP_NM_NUM	브랜드 주택 여부 (binaray)
HUS_TP_아파트	아파트 여부 (binaray)
HUS_TP_오피스텔	오피스텔 여부 (binaray)
HUS_TP_연립다세대	연립다세대 여부 (binaray)
log_price	로그변환 매매가

$$RMSLE = \sqrt{(log(y_i + 1) - log(\hat{y}_i + 1))^2}$$

Fig. 1. RMSLE

참고문헌[6]에 따라 평가지표를 산출할 머신러닝 비교모델은 일반적인 회귀 모델 4가지와 결정트리인 CART (Classification And Regression Tree) 알고리즘의 4가지 모델을 비교한다. 각 모델의 특징을 Table 2에 정리한다. 참고문헌[7]에 따르면, 기존 연구들은 주로 ARIMA 모형과 딥러닝 기반의 DNN 이나 LSTM 모델을 사용하였으나, 본 연구는 머신러닝 기법 중 여러 회귀 모델과 CART 모델을 비교하여 사용한다는 점에서 차별성을 갖는다.

모델의 정확도를 높이기 위해 교차검증을 진행하는데, 일반적인 k-fold 교차검증으로 사용하지 않고, 시계열 데이터 분석에서 자주 사용되는 TimeSeries 교차검증 기법을 적용하였다. k-fold 교차검증은 k개의 fold를 구성 후 k-1개의 fold를 학습 데이터세트로, 나머지 1개의 fold를 검증 데이터세트로 사용하여, 과적합을 방지하고 좀 더 일반화된 모델을 만들 수 있다. 하지만 이 연구를 위해 구성한 데이터세트에 적용 시, 미래 데이터를 통해 과거의 데이터를 예측하는 오류를 범할 수 있다.

Table 2. Usage model and characteristics of the model

사용모델	특징
Linear Regression	설명변수와 종속변수의 상관관계를 반영한 일반 선형회귀 학습모델
Ridge Regression	L1 규제를 가진 선형회귀 학습모델
Lasso Regression	L2 규제를 가진 선형회귀 학습모델
ElasticNet Regression	L1&L2 규제를 결합한 선형회귀 학습모델
Decision Tree	트리의 불 순도가 낮은 방향으로 분기하며 불 순도를 최소화하는 방향으로 학습하는 트리 기반 학습모델
Random Forest	Bagging을 사용, 랜덤하게 변수 선택함으로써 Decision Tree의 과적합을 방지한 트리 기반 학습모델
XGBoost	Ensemble, Boosting 기법을 활용하여 이전 모델의 Loss를 학습데이터에 입력, Gradient 방법을 이용해 오류를 보완한 트리 기반 학습모델
LightGBM	GOSS, EFB, Leaf Wise 방법을 활용한 오류 손실 최소화한 트리 기반 학습모델

그에 비해, Time Series Nested Cross Validation 기법[8]은 전후 상관관계가 있는 데이터일 때 사용하는데, 훈련 데이터세트를 검증데이터 세트보다 항상 앞선 시간에 할당하는 기법이다. 그래서 전후 상관관계가 있는 데이터세트 특성상 일반적인 k-fold 기법을 사용하지 않고, Time Series Nested Cross Validation 기법을 사용한다. 사용언어는 파이선을 사용하고 사이킷런 라이브러리의 Time Series Split를 사용한다.

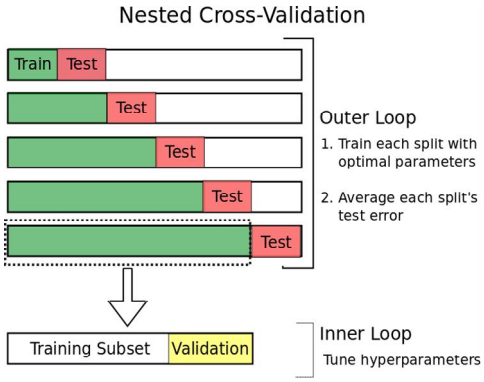

Fig. 2. Time Series Nested Cross-Validation

Fold 개수는 시스템 성능을 고려하여 10개로 구성한다. 비교모델 8개에 대하여 각각 10개의 fold를 교차검증한 결과를 내어 평균값을 산출한 뒤, 가장 낮은 평균 RMSLE를 갖는 모델을 선택한다. Fig. 3에 제시된 결과처럼 전반적으로 CART 기반 모델들이 예측력이 좋은 것으로 확인할 수 있다. 그 중 LightGBM의 RMSLE 0.24821로 가장 좋은 정확도를 보이기 때문에 LightGBM를 본 연구를 위한 학습모델로 채택한다.

	Model	Average RMSLE Score
0	LinearRegression	0.502020
1	Ridge	0.501945
2	Lasso	0.502551
3	ElasticNet	0.498835
4	DecisionTreeRegressor	0.414556
5	RandomForestRegressor	0.336892
6	XGBRegressor	0.258437
7	LGBMRegressor	0.248216

Fig. 3. Usage model Average RMSLE Score

LightGBM 모델의 정확도를 향상시키기 위해 하이퍼파라미터 조정이 필요하다. 하이퍼파라미터 선택은 OPTUNA 프레임워크[9]를 사용하는데, OPTUNA는 각 하이퍼파라미터에 범위를 할당하면 범위 내에서 목적에 맞게 여러 번 평가지표를 확인하여 최적화한다. 가장 적합한 하이퍼파라미터를 찾아 선택하는 프레임워크이다. GridSearchCV와 비교하여 범위를 할당하면, 자동으로 최적값을 찾아주는 장점 때문에 사용자 편의적인 측면에서 더 부합한다. RMSLE값이 낮을수록 높은 정확성을 갖기 때문에, direction = 'minimize'로 설정하고, 시스템 성능을 고려하여 반복 횟수는 n_trial = 100으로 설정한다. LightGBM 모델의 RMSLE 값의 결과는 0.22312로 하이퍼파라미터 조정 전의 값 0.24821보다 약 10.2% 정확도가 향상된 것을 알 수 있다.

모델에 적용한 주택 매매 데이터세트의 특성 중 어떤 특성들이 높은 영향력을 끼치는지 파악하기 위해 특성 중요도 분석을 수행하였으며 Fig.4에 중요도에 따른 결과 그래프를 보인다.

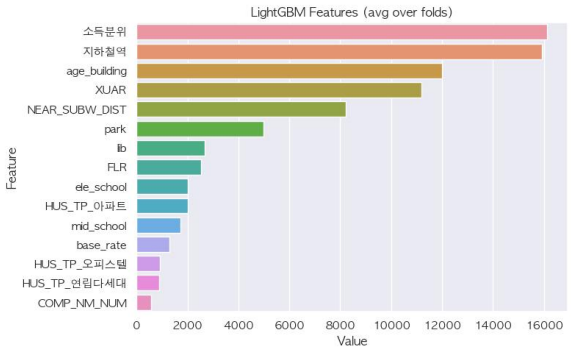

Fig. 4. Result of feature importance

주택가격 형성의 가장 큰 영향을 가진 요인은 주택이 위치한 지역의 평균 소득분위인 것을 알 수 있으며, 두 번째 요인은 주변에 위치한 지하철역이 이용량이 많은 주요 지하철역일 때 영향도가 큰 것을 알 수 있다. 세 번째 요인은 주택 내부요인인 건물의 오래된 정도가 영향이 큰 것으로 알 수 있다. 외부요인인 위치를 통해 도출된 특성과 위치에 기반한 생활-SOC 요인이 내부요인보다 주택매매 가격 형성에 더 큰 영향을 준다는 것을 알 수 있다.

III. Conclusions

본 논문은 주택 매도, 매수 시에 주로 확인하는 요인들을 활용하여 주택가격 예측 및 영향 요인 분석 목적으로 연구를 진행하였다. 소득분위가 높은 지역에 따라 집값의 영향력이 높았고, 그다음으로는 주요 지하철역일수록 영향력이 높았다. 해당 지역의 평균 소득분위가 높고 주변에 주요 지하철역이 위치할 때 높은 주택가격대를 형성한다는 것을 알 수 있다. 교통 부분 사회기반시설 특성과 지역의 평균 소득분위 특성의 상관계수가 +0.4인 점을 감안하면, 대부분의 소득분위가 높은 지역에서 교통도 좋은 편으로 보인다. 그러므로 사회기반시설 부분에서는 교통 부분 사회기반시설이 가장 중요하다고 판단된다. 추후 다양한 영향요인을 추가하면 더 좋은 모델을 구성할 수 있다고 판단하여, 주택 매수 시 중요 요인을 해당 모델에 추가 입력하여 연구하고자 한다. 향후 주택 매수, 매도 시에 적정가격을 판단하는 지표로 활용되길 기대한다.

ACKNOWLEDGEMENT

이 논문은 2021년도 정부 (과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 이공분야기초연구사업임 (NRF-2021R1F1A1064073).

REFERENCES

[1] Korea Real Estate Board station area actual transaction data, https://www.bigdata-transportation.kr/frn/prdt/detail?prdtId=PRDTNUM_000000020052

[2] Seong Wan Bae. “Forecasting Property Prices Using the Machine Learning Methods:Model Comparisons.” Dissertation, Dankook University, 2019.

[3] Seoul Transportation Corporation Urban Railway History Information, https://data.seoul.go.kr/dataList/OA-15442/S/1/datasetView.do

[4] Incheon Transit Corporation Urban Railway History Information, https://www.data.go.kr/data/15083751/fileData.do?recommendDataYn=Y

[5] T-money transportation card statistical data, https://www.t-money.co.kr/ncs/pct/ugd/ReadTrcrStstList.dev

[6] OPTUNA, https://dacon.io/codeshare/2704

[7] Seong Wan Bae, Jung Suk Yu. “Predicting the Real Estate Price Index Using Deep Learning” Korea Real Estate Research Institute: Korea Real Estate Review 27(3), 2017, pp. 71-86.

[8] Cochrane, Courtney. “Time Series Nested Cross-Validation.” Towardsdatascience, Towards Data Science, 19 May 2018, https://towardsdatascience.com/time-series-nested-cross-validation-76adba623eb9.

[9] DACON, https://dacon.io/competitions/official/21265