韓國地籍學會誌 第39卷 第3號 2023. 12(3) : 199~210 https://doi.org/10.22988/ksc.2023.39.3.014

머신러닝을 활용한 부동산 실거래가 요인 분석

Real Estate Transaction Price Factor Analysis Using Machine Learning

박 서 현* · 김 도 형**
Park, Seo Hyeon · Kim, Do Hyoung

요 약

본 연구에서는 사람들의 거주 및 업무 등에 가장 밀접한 요소 중 하나인 부동산 실거래가가 어떠한 요인에 의하여 높아지고 낮아지는지를 파악하고자 하기 위해 토지, 인구, 업종에 대한 데이터들을 활용하여 요인 분석을 진행하였다. 연구를 원활하게 진행하기 위해 공간적 범위는 대구광역시 법정동을 기준으로 설정하였으며, 시간적 범위는 2020년 7월부터 2023년 6월까지의 기간을 설정하였다. 요인 분석을 진행하기 전에 실거래가에 대한 분포가 어떻게 구성되어 있는지 파악하기 위해 KDE plot과 Box plot을 사용하여 기초통계를 확인한 결과 특정 구간에 밀집되어 있으며 오른쪽으로 긴 형태인 Positive skewness 형태가 보였다. 이러한 실거래가의 분포를 활용하여 각각의 독립변수의 영향 정도를 알아보기 위해 실거래가 분포를 기준으로 4개의 Group으로 나눈 후 속도 및 예측력에서 뛰어난 성능을 보이는 머신러닝의 한 종류인 XGBoost를 활용하여 독립변수에 관한 영향도 분석을 진행하였다. 분석 결과 부동산 실거래가에 미치는 영향은 Group별로 상이함을 확인하였고, 공간적 분포에 따라 실거래가 및 독립변수의 두드러진 차이를 확인할 수 있었다. 향후 연구에서는 이러한 구간별 독립변수의 영향력과 공간적 분포를 바탕으로 실거래가를 예측하는데 중요한 정보가 될 수 있을 것으로 예상된다.

주요어 : 머신러닝, 커널밀도함수, XGBoost, 부동산, 실거래가

ABSTRACT

In this study, factor analysis was conducted using data on land, population, and industry to understand what factors increase and decrease the actual real estate transaction price, which is one of the most close factors to people's residence and work. In order to facilitate the study, the spatial range was set based on jurisdiction, Daegu, and the temporal range was set from July 2020 to June 2023. The dependent variable of the data used in the analysis is the actual transaction price, and the independent variable consists of a total of 70 columns, including the land category, use area, demographics, and number of businesses. Before conducting factor analysis, basic statistics were checked using KDE plot and Box plot to understand how the distribution of actual transaction prices was structured, and the actual transaction price was concentrated in a specific section and the shape of a positive skewness with a long tail to the right was confirmed. To determine the degree of influence of each independent variable, the actual transaction price was divided into four groups

* 주저자, 정회원·성동구청 정보통신과 빅데이터센터 주무관(E-mail: psh0196@uos.ac.kr)
** 교신저자, 정회원·경일대학교 부동산지적학과 조교수(E-mail: do@kiu.ac.kr)

according to the distribution, and the effect on the independent variable was analyzed using XGBoost, a type of machine learning with fast speed and predictive performance. As a result of the analysis, the effect on the real estate transaction price was different for each group, and it was confirmed that the actual transaction price or the value of the independent variable had a significant difference according to the spatial distribution. Future research is expected to be important information for predicting actual transaction prices based on the influence and spatial distribution of independent variables by group.

Keywords : Machine Learning, Kernel Density Estimation, XGboost, Real Estate, Transaction Price

1. 서 론

1.1 연구의 배경 및 필요성

부동산 매매가격은 경제·사회·정책적 상황에 따라서 변동되는데, 부동산의 매매 시에 가장 중요한 요인 중 하나가 바로 수요자의 효용이다. 수요자는 해당 부동산이 지닌 위치, 크기, 주변 환경 등의 특성을 파악하여 생활을 영위하기 위한 필요와 욕구를 만족시켜야 해당 부동산을 가치 있는 것으로 판단하고 해당 부동산 거래를 수행하고자 한다. 만족하는 부동산을 취득하고자 할 때의 가격에 관한 참고자료는 정부에서 공개하고 있는 공시지가와 실거래가 등을 활용하고 적정 가격을 설정한다. 국토교통부는 감정평가를 통해 매년 1월 1일 부동산에 관한 공시지가를 공시하고 이를 통해 세금, 보상금, 거래가격 등의 기준을 설정하고 있으며, 부동산 실거래가는 2006년 1월부터 이중계약 등의 잘못된 관행을 없애고 부동산 거래를 투명하게 하기위해 부동산거래 신고 및 주택거래 신고를 한 아파트, 연립주택, 다세대주택, 단독주택, 다가구주택, 토지, 창고 등을 대상으로 부동산 실거래가 공개 제도를 도입하면서 시작되었다. 또한, 최근 4차 산업혁명 시대는 초연결, 초지능 등의 메가트렌드가 발생하였고, 더불어 4차 산업혁명 시대의 핵심 기술인 인공지능 분야가 대두 시작하였다. 인공지능은 군사, 환경, 레저, 공간정보 등의 분야에서 다양하게 활용하고 있다. 특히, 머신러닝은 데이터를 기반으로 일정한 현상의 패턴 등을 학습하여 성능향상 및 결과 분석, 예측 등을 지원하고, 머신러닝의 결과는 사용자의 의사결정에 도움을 준다. 이러한 머신러닝 기술을 적용하여 부동산 실거래가에 미치는 복잡한 주변 환경의 공통적인 요인들을 모색하고 향후 부동산 실거래가 예측에 있어 독립변수의 영향력과 공간적 분포가 중요한 정보가 될 수 있음을 확인하고자 한다.

1.2 연구의 목적 및 방법

부동산 매매가격은 정책적인 부분으로 인하여 많은 사람들의 관심과 연구 등을 통해서도 예측하기 어려운 부분이 존재하지만, 부동산 매매가격은 해당 부동산의 주변 요인들에 따라서 변동되기도 하여 해당 부동산의 주변 요인들에 따라 부동산의 실거래 가격에 차이가 날 수 있다고 예측할 수 있다. 이에 본 연구에서는 기존의 부동산 실거래가 통계자료를 토대로 해당 부동산 주변의 지목, 인구, 상업 등의 요인들을 토대로 머신러닝을 이용하여 부동산 실거래가 요인을 분석하고자 한다.

연구의 원활한 진행을 위해 부동산매매가 활발히 진행되고 있는 대구광역시의 법정동을 기준으로 진행하였다. 연구의 효율적인 성과를 나타내기 위해 부동산가격의 변동률이 뚜렷하게 나타나고, 면적이 다른 지역에 비하여 비교적 넓은 지역인

대구광역시를 연구지역으로 선정하였다. 데이터의 기간은 2020년 하반기부터 2023년 상반기까지 최근 3년 동안의 부동산 실거래가를 기준으로 설정하였다. 연구방법은 예측에 뛰어난 성능을 보이며 각 독립변수가 예측값에 미치는 영향요인을 파악할 수 있는 머신러닝의 한 종류인 XGBoost에 적용하기 위해 국토교통부 및 행정안전부 등의 부동산에 관한 자료, 인구에 관한 자료 등의 통계자료를 수집하여 날짜와 법정동을 기준으로 데이터 전처리를 수행하였다.

1.3 선행연구 고찰 및 차별성

배성완·유정석(2017)은 딥러닝 방법을 이용하여 부동산가격지수 예측에 적용하여, 기존의 시계열 분석 방법과의 비교를 통해 부동산 시장 예측에서 활용 가능성을 확인하고자 하였다. 이를 위해 딥러닝 방법인 DNN모형 및 LSTM모형을 이용하여 여러 가지 부동산가격지수에 대한 예측을 시행하였다. 시행 결과 딥러닝 방법이 시계열 방법보다 우수하였고 DNN모형이 LSTM모형보다 미미하지만 예측력이 우수한 결과를 도출하였다. 이에 딥러닝 방법을 활용하여 부동산 시장에 대한 예측의 정확성을 제고 할 수 있을 것으로 판단하였다.[1] 고주형·강명구(2019)는 서울시 재건축 완료 아파트의 가격과 가격상승률에 영향을 미치는 주거특성 요인을 비교·분석하여 향후 아파트 가격 상승률의 예측을 위한 기초자료로 활용하고자 하였다. 주거특성 요인에 따른 아파트 가격요인 및 가격상승률에 대하여 선형모형으로는 영향의 크기를, 로그-로그모형으로는 가격형성의 요인별 탄력성을 추정하였다. 연구 결과, 아파트 가격과 아파트 가격 상승률에 공통으로 영향을 미치는 요인은 지하철역 개수, 과거시세, 시차변수이며 차이점은 강남 4구 지역성 변수로 도출하였다.[2]

송기욱·류강민(2019)은 산업용 부동산의 대표적 투자자산 유형으로서 물류부동산의 실거래가에 영향을 미치는 결정요인을 실증 규명하고자 하였다. 이를 위해 수집된 분석 대상의 개별적 특성요인들과 실거래가의 관계를 검증하고자 헤도닉가격모형을 사용하였고, 투입변수를 달리해 3가지 모형으로 나누어 추정하였다. 분석결과, 물류부동산은 운송비용 절감이 가능한 곳에 불특정하게 집중분포하는 경향을 보였고, 입지, 건물, 토지, 시간을 포함한 개별적 특성요인들이 실거래가와 밀접한 관련성을 지닌 것으로 파악하였다. 이에 물류부동산의 실거래가 변동을 예측하는데 기초자료로 활용하고자 하였다.[3] 김학현 외 2인(2022)은 다양한 부동산 사이트에서 자료 수집 및 크롤링을 통해 2015년부터 2020년까지 87만개의 방대한 데이터셋을 구축하고 다양한 아파트 정보와 경제지표 등 가능한 많은 변수를 모은 뒤 미래 아파트 매매실거래가격을 예측하는 모델을 구축하였다. 또한, 심층신경망(DNN), XGBoost, CatBoost, Linear Regression 총 4개의 머신러닝 및 딥러닝 알고리즘을 이용해 하이퍼파라미터 최적화 후 모델을 학습시키고 모형 간 예측력을 비교하였다. 이후 실제 2021년 데이터와 비교한 결과 예측가능한 성과를 만들었으며, 이를 통해 머신러닝과 딥러닝이 아파트 실거래가 예측을 할 수 있을 것으로 판단하였다.[4]

선행연구는 딥러닝을 이용하여 부동산가격지수에 관한 예측, 아파트 가격 상승률에 영향을 미치는 요인 비교·분석, 헤도닉가격모형을 활용하여

1) 배성완·유정석, “딥 러닝을 이용한 부동산가격지수 예측”, 「부동산연구」, 제27집 3호, 2017, pp.17-86.
2) 고주형·강명구, “부동산 가격 요인과 가격상승률 요인 비교 연구:서울시 재건축 아파트를 중심으로”, 「부동산학연구」, 제25집 2호, 2019, pp.7-22.
3) 송기욱·류강민, “헤도닉가격모형을 이용한 물류부동산의 실거래가 결정요인 실증분석”, 「부동산학연구」, 제25집 3호, 2019, pp.23-37.
4) 김학현·유환규·오하영, “딥러닝과 머신러닝을 이용한 아파트 실거래가 예측”, 「정보처리학회논문지 소프트웨어 및 데이터 공학」, 제12권 2호, 2023, pp.59-76.

물류부동산에 관한 실거래가 요인 분석, 머신러닝을 활용하여 아파트 실거래가 예측 분석 등을 수행하였다. 본 연구는 누구나 쉽게 구할 수 있는 정부에서 제공되고 있는 공공데이터인 지목, 토지이용현황, 인구, 상업입지 등을 이용하고 머신러닝을 활용하여 부동산실거래가에 영향을 미치는 요인을 분석하였다. 이후 부동산실거래가에 영향을 미치는 요인들의 주변 환경을 통해 부동산실거래가를 예측하고 이후 부동산 거래와 정책적으로 필요한 의사결정을 하는데 중요한 기초자료로 활용될 수 있도록 한다는 점에서 그 차별성이 있다.

2. 이론적 고찰

2.1 탐색적 데이터 분석

탐색적 데이터 분석(Exploratory Data Analysis, EDA, 이하 EDA)는 수집된 데이터들을 다양한 각도에서 관찰하고 이해하는 과정을 말하며, 데이터가 표현하는 현상을 도표 등의 수치와 그래프 등의 그림 형태로 표현한다. Tukey(1979)[5]는 탐색적 데이터 분석을 통해 데이터가 표현하는 현상의 이해, 데이터의 수집이나 가공 여부에 대한 결정, 기존에 알 수 없었던 패턴의 발견, 이를 통해 새로운 가설을 만들 수 있다는 점에서 매우 중요하다고 주장했다. 다양한 분야의 데이터로부터 정보를 얻어내기 위해 데이터에 대한 이해가 필요하다는 것을 의미한다고 볼 수 있다.[6]

EDA의 방법론 중 하나의 커널밀도 추정방법(Kernel Density Estimation, KDE)은 관측 데이터가 가진 확률 밀도를 표현하는 방법으로 시공간적 분포가 다양하게 관측되는 현상에 유연적인 활용도가 높다. 1차원일 때에는 대표적으로 히스토그램과 같은 형태로 나타나고, 2차원은 등고선의 형태로 데이터의 밀집도를 확인하여 Hotspot 분석 등을 진행할 수 있다.[7]

다양한 데이터로부터 적절한 의미도출을 위해서는 EDA가 필수적으로 요구되므로 본 연구에서는 수집된 데이터들의 각각의 의미를 도출하기 위해 커널밀도함수의 가장 기본적인 형태와 Box plot을 활용하여 종속변수인 실거래가의 분포를 파악하고자 한다.

2.2 XGBoost (Extreme Gradient Boosting)

본 연구에서는 지목, 인구, 사업체, 용도지역 등의 독립변수의 영향력과 상관관계를 파악하기 위해 머신러닝 모델 중 성능이 뛰어난 트리모델인 XGBoost를 활용하였다. Boosting기법 중 가장 널리 활용되는 알고리즘으로 기저모형[8]들을 순차적으로 구축한 후 최종적으로 병합(가중평균)하여 의사결정을 진행한다. 특히, 전 단계에서 잘못 분류된 개체에 더 많은 가중치를 주고 모형을 구축하기 때문에 정확도와 속도가 모두 뛰어나 현재 다양한 예측 및 분류 문제에 많이 활용되고 있다.[9] XGBoost의 예측 성능을 활용하여 본 연구에서는 실거래가가 어떠한 독립변수에 의해 높아지고 낮아졌는지를 파악한다. 대표적으로 두 가지의 방법을 활용하는데, 먼저 예측에 있어 각 변수의 영향력을 파악하는 변수중요도(feature importance)가 있다. XGBoost에서 변수중요도를 산정하는 방법은 변수별 데이터를 분리하도록 활용한 가

5) Tukey, “Exploratary Data Analysis”, Journal of the Royal Statistic Society, Vol. 28(1), 1979, pp.79-83.
6) 강성경, “증거기반정책을 위한 행정자료의 탐색적 데이터 분석(EDA)과 활용 : 선박안전 분야 관련 속성 분석 및 정책 적용을 중점으로”, 박사학위논문, 동국대학교 대학원, 2021, pp.14-16.
7) M Kalinic, 2018. Kernel Density Estimation (KDE) vs. Hot-Spot Analysis - Detecting Criminal Hot Spots in the City of San Francisco.
8) 서로 직교하면서 선형적으로 독립적인 함수의 집합을 의미.
9) https://ko.wikipedia.org/wiki/XGBoost

중치(Weight), 해당 변수로 분리된 데이터의 수(Cover), 변수를 사용했을 때 오차가 줄어드는 평균적인 train loss(Gain) 이 3가지 중 한가지 지표를 활용하여 각각의 변수가 추가되었을 때 영향력을 산정하며 디폴트는 Grain으로 많이 활용한다. 하지만 변수중요도는 어떠한 영향(음의 영향 또는 양의 영향)을 미치는지 판단을 할 수 없고 변동성이 있으므로, 이를 보완하기 위해 Shap Value를 사용한다. Shap Value는 각각의 변수에 대하여 Sampling 후 실제값과 예측값의 차이를 계산하고, 반복적인 결과의 평균을 낸 수치이기 때문에 변동성이 거의 없을뿐더러 양의 영향력인 것인지 음의 영향력인지를 판단해주는 지표가 된다. XGBoost 뿐만 아니라 다른 모델에서도 Shap value, feature importance를 제공하고 있으며, 대표적으로 Random -forest와 XGBoost 등이 있다. 이중 XGBoost는 CPU캐시를 고려한 알고리즘으로 데이터의 용량이 커지더라도 빠른 속도로 학습할 수 있다는 장점이 있어 본 모델을 활용하였다.

3. 부동산실거래가 요인 분석

3.1 데이터 수집 및 전처리

실거래가의 거래가격 요인 분석을 위해 대구광역시의 법정동을 기준으로 2020년 7월부터 2023년 6월까지 약 3년간의 데이터를 70개의 칼럼으로 구성하였다. 행 개수는 총 3,050개로 구성되어 있으며 모든 칼럼을 분석하기 위해 기준년 월, 법정동 코드를 Key 값으로 활용하였다. 종속변수는 국토교통부의 실거래가(만원/㎡)[10]이며, 독립변수는 실제 공시지가의 설정기준에 포함된 토지의 특성과 그 외에 실거래가의 추가적인 요인일 것으로 예상되는 인구, 업종의 특성을 반영하기 위하여 국토교통부의 지목(면적의 합계)[11], 용도지역(면적의 합계)[12], 행정안전부의 인구통계(명)[13], 소상공인시장진흥공단의 사업체 수(업종별 개수)[14]로 Key 값을 기준으로 데이터를 조인하여 분석에 활용될 최종 데이터를 구성하였다.

〈표 1〉 데이터 테이블

no	구분	칼럼명	시간범위(단위)	공간범위(단위)
1	key	기준년월	2020.7.~2023.6. (월별)	-
2		법정동코드	-	대구광역시(법정동)
3	실거래가	실거래가	2020.7.~2023.6. (월별)	대구광역시(법정동)
4	지목	공원	2020.~ 2023. (연도별)	
5		공장용지		
6		과수원		
7		광천지		
8		답		
9		대		
10		도로		
11		목장용지		
12		유원지		
13		유지		

10) 국토교통부, https://rt.molit.go.kr/
11) 국토교통부, https://www.data.go.kr/data/15004246/fileData.do
12) 국토교통부, https://www.data.go.kr/data/15004246/fileData.do
13) 행정안전부, https://jumin.mois.go.kr/
14) 소상공인시장진흥공단, https://www.data.go.kr/data/15083033/fileData.do

14		임야		
15		잡종지		
16		전		
17		종교용지		
18		주유소용지		
19		주차장		
20		창고용지		
21		철도용지		
22		체육용지		
23		하천		
24		학교용지		
25	용도지역	개발제한구역	2020.~ 2023. (연도별)	대구광역시(법정동)
26		근린상업지역		
27		농림지역		
28		보전녹지지역		
29		생산녹지지역		
30		유통상업지역		
31		일반공업지역		
32		일반상업지역		
33		자연녹지지역		
34		자연환경보전지역		
35		제1종일반주거지역		
36		제1종전용주거지역		
37		제2종일반주거지역		
38		제2종전용주거지역		
39		제3종일반주거지역		
40		준공업지역		
41		준주거지역		
42		중심상업지역		
43	인구통계	총거주자수	2020.7.~2023.6. (월별)	대구광역시(법정동)
44		0~9세		
45		10~19세		
46		20~29세		
47		30~39세		
48		40~49세		
49		50~59세		
50		60~69세		
51		70~79세		
52		80~89세		
53		90~99세		
54		100세이상		
55		남성 거주자		
56		여성거주자		
57	사업체 수	과학·기술	2020.7.~2023.6. (분기별)	대구광역시(법정동)
58		관광/여가		

59		교육		
60		보건의료		
61		부동산		
62		생활서비스		
63		소매		
64		수리·개인		
65		숙박		
66		스포츠		
67		시설관리·임대		
68		예술·스포츠		
69		음식		
70		학문/교육		

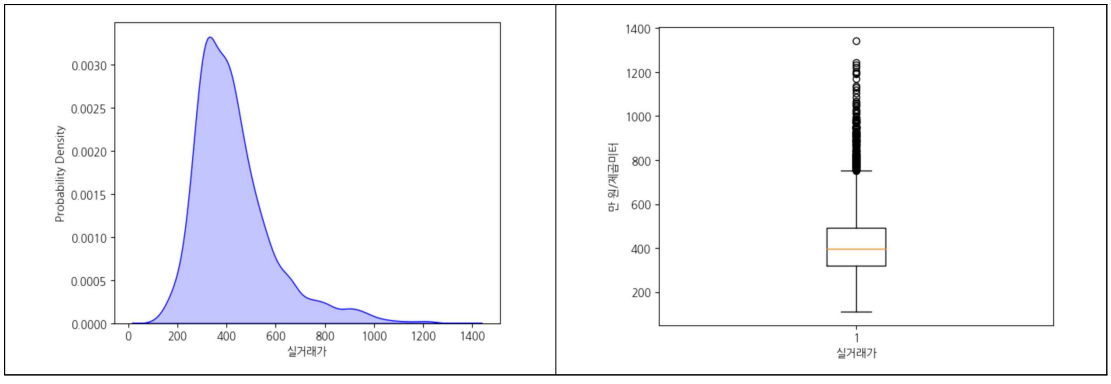
3.2 실거래가 통계분포

[그림 1]은 대구광역시의 실거래가의 KED Plot과 Box Plot을 나타낸 것으로 대구광역시의 실거래가 데이터를 분석한 결과 ㎡당 200만 원에서 600만 원의 가격이 가장 많이 분포되어 있었으며, 평균적으로 428.3만 원으로 나타났다. 데이터의 밀집과 연속변수의 특성으로 인하여 일반적으로 사용하는 통계적 분위수를 활용하여 Group을 구분하였다. 1사분위수(25%)는 318.87만 원, 2사분위수(50%)는 396.02만 원, 3사분위수(75%)는 492.31만 원으로 각각의 구간에 따라 관련된 요인분석을 진행하였다.

Group1(0~1사분위수)은 763개, Group2(1사분위수~2사분위수)는 762개, Group3(2사분위수~3사분위수)는 762개, 4Group(3사분위수~max)는 763개로 나타났는데 실거래가의 분포는 평균과 중위수를 비교하였을 때 outlier가 높은 값으로 많이 분포되어 있어 평균이 더 크고 왼쪽으로 밀도가 치우친 형태인 Positive skewness로 볼 수 있다.

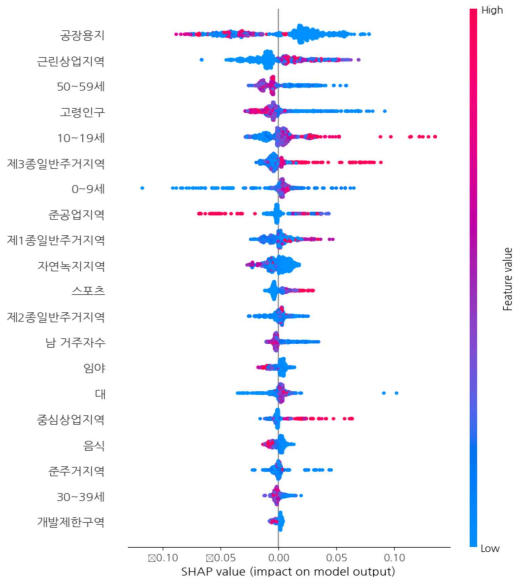
3.3 실거래가 요인 분석 결과

본 연구에서 사용한 XGBoost는 python의 sckit-learn이라는 내장 모듈을 활용하였다. 하이퍼파라미터는 본 모델을 휴리스틱한 방법으로 변경해가며 조절하였으며 최종적으로 의사결정 나무의 개수를 의미하는 n_estimators는 100, 가중치에 대해 구해진 기울기 값을 얼마나 경사하강법



〔그림 1〕 실거래가 분포 KDE Plot(좌)과 Box Plot(우)

에 적용할지 결정하는 학습률 learning rate는 0.08, 트리의 최대 깊이를 산정하는 파라미터인 max depth는 7로 구성하였다. 이론적 고찰에 나타나는 XGBoost의 작동방법, Shap value, 변수중요도 산정방법을 바탕으로 구현한 결과 전체 실거래가에 대한 Shap value 결과는 [그림 2]와 같이 나타났다.



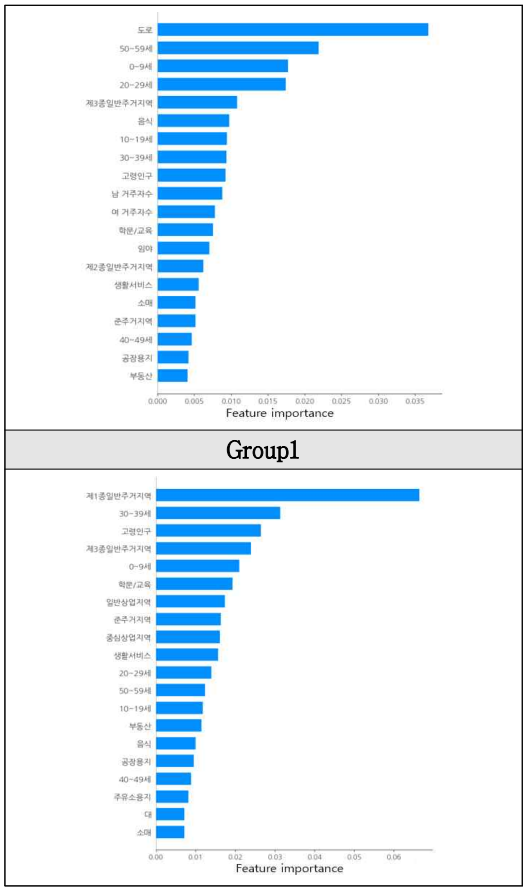
〔그림 2〕 독립변수별 Shap Value 산정 결과

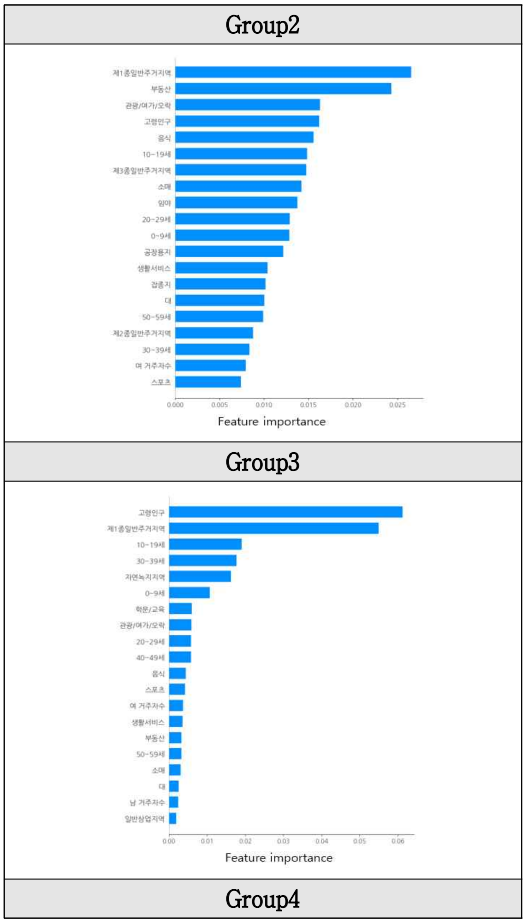
Shap value는 X축 중심위치 0을 기준으로 오른편으로 치우칠수록 예측에 있어서 양(+)의 영향력을 미치고 왼편으로 갈수록 음(-)의 영향력을 미친 것으로 해석되고 길이가 길어질수록 높은 영향력을 미친다는 것을 의미한다. [그림 2]와 같이 공장용지의 경우 값이 작을수록 예측값이 크게 나타나는 경향이 있고, 값이 높을수록 예측값을 작게 예측하는 경향이 있다고 볼 수 있다. 왼편과 오른편을 비교하였을 때, 왼편이 더 긴 것을 보아 음의 영향력이 더 큰 것이다. 위에서 아래로 나열된 변수의 순서는 전체적으로 영향을 많이 끼치는 순으로 독립변수를 나열한 것이라고 할 수 있다.

앞서 설명된 해석방법을 바탕으로 실거래가와 전체의 독립변수 요인을 보았을 때 공장용지가 넓을수록 실거래가가 낮아지고 좁을수록 실거래가가 높아지는 것을 볼 수 있다. 이에 반해 근린상업지역은 넓을수록 실거래가가 높아지는 현상을 나타내고 있다. 근린상업지역 아래의 고령인구, 학령인구 등의 영향을 볼 수 있는데 이것은 학령인구는 양의 상관관계를 가지고 있고 고령인구는 음의 상관관계로 나타났다. 대부분 토지의 용도 또는 인구가 실거래가에 영향력이 크게 나타났으며 전체 실거래가에 대하여 업종은 영향력 정도가 낮지만, Group에 따라 업종의 영향력이 크게 나타나는 현상을 볼 수 있다.

다음으로 실거래가의 시공간적 분포에 따라 각각의 독립변수의 영향 정도 차이를 알아보기 위해 실거래가 기준으로 4개의 Group을 분할 하여 독립변수별 영향도를 확인하여 보았다<표 2>.

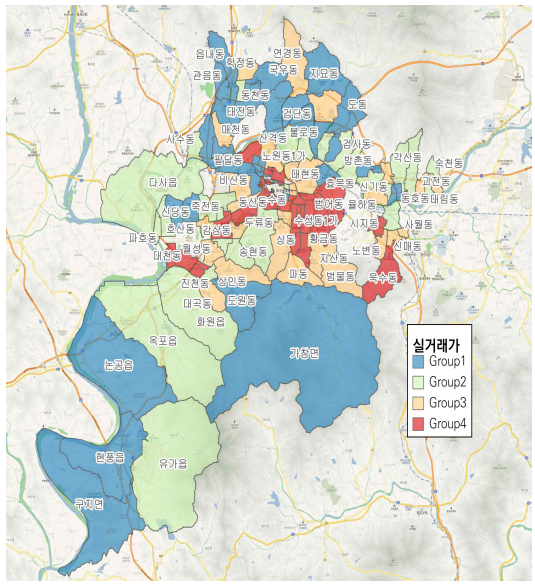
〈표 2〉 Group별 변수 중요도





영향도 확인 결과, Group1에서는 도로의 면적의 영향력이 가장 높았으며 읍면동에서 도로 지목의 면적이 크게 나타난 것이 실거래가에 영향을 끼친 것으로 보이며, 다음으로는 인구와 제3종 주거지역 순으로 나타났다. Group2에서는 제1종주거지역이 가장 큰 것으로 나타났으며 경제활동 인구 중 하나인 30대, 고령인구, 제3종일반주거지역 등이 변수중요도에서 높게 나타났으며 Group1과 비교하였을 때 학문/교육과 일반상업지역, 생활서비스 등이 더 높게 나타난 것을 볼 수 있다. Group3에서는 다른 Group과 다르게 업종의 영향력이 크게 나타났다. 제1종일반주거지역의 영향력이 가장 높지만, 다음으로 부동산업, 관광/여가/오락이 나타났다. 그 뒤로 고령인구, 음식업, 학령인구 순으로 나타났다. Group4에서는 고령인구와 제1종일반주거지역의 변수중요도가 압도적으로 높았다. 이렇듯 실거래가의 분위수별로 변수중요도가 다르게 나타났는데, Group3에서 비교적 업종이 실거래가에 미치는 영향력이 컸고, 실거래가가 가장 높은 Group인 Group4의 경우는 고령인구 수 또는 주거지역 용도의 특성이 실거래가에 큰 영향을 준 것으로 보인다.

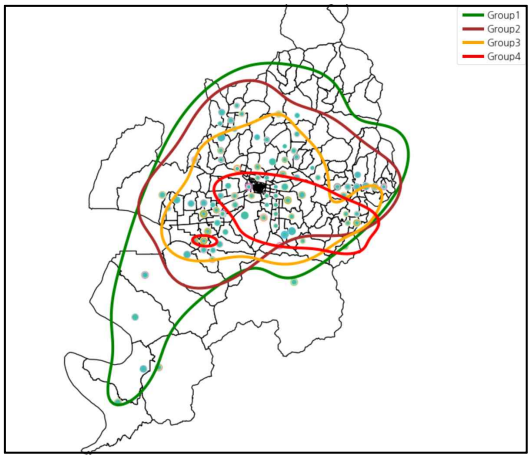
다음으로 [그림 3]은 요인별 분석 결과를 바탕으로 분석 기간의 실거래가 평균을 Group별로 나눈 결과이다. 시각화가 되지 않은 부분은 실거래가가 나타나지 않은 지역이거나, 결측값이 있는 지역으로 판단된다.



〔그림 3〕 대구시 Group별 실거래가 공간적 분포

공간적 분포를 살펴본 결과 도심을 중심으로 실거래가가 높은 Group이 분포하고 있고 외곽으로 갈수록 실거래가가 낮아지는 패턴을 나타냈다. Group1의 경우 도로의 영향이 많았는데 구자읍, 현풍읍, 읍내동, 관음동이 비교적 고속도로의 면적이 크게 나타난 것으로 보인다. Group2는 제1종일반주거지역의 변수 영향도가 가장 큰 것으로 나타났었는데, 처음 전체변수의 Shap value의 결과에서 제1종일반주거지역은 양의 상관관계가 나타났

다. 따라서 그림에 나타난 Group2는 제1종일반주거지역의 면적분포에 따라 실거래가에 양의 상관관계가 많이 미치는 지역으로 해석할 수 있다. Group3에서는 다른 Group에 비해 업종의 영향력이 상위권으로 보였는데, 그림의 Group3에서는 실거래가가 상권의 영향과 근무를 목적으로 근처 거주를 하는 사람들이 실거래가에 영향을 끼쳤다고 예상할 수 있다. 마지막으로 국토연구원(2022)[15]은 대구의 노인인구 변화율의 증가속도가 빠르고, 총인구변화율은 감소한다고 하였다. 그러한 특징이 부동산 실거래가에도 노인인구의 영향이 Group4뿐만 아니라 전체 실거래가의 값 영향력에서도 높게 나타난 것을 확인하였다.



〔그림 4〕 Group별 독립변수 KDE plot

마지막으로 각각의 행정동을 중심으로 독립변수들의 밀도를 파악하기 위해 Group1~4에서 각각 산출된 변수중요도를 커널밀도함수를 활용하여 지도에 표현하였다[그림 4]. 단순히 행정동 중심점을 기준으로 커널밀도함수를 적용하기보다 인구의 생활반경을 기준으로 독립변수의 밀집도를 파악하기 위해 행정동 내에서의 인구밀도 무게중심점을 산출하였다. 국토정보플랫폼에서 제공하는 1Km 격자별 총인구수 데이터를 활용하여 각각의 행정동별로 각각의 중심점을 산출하고, 산출된 중심점을 기준으로 변수중요도에 나타난 Top 20의 독립변수들을 지도에 커널밀도함수로써 시각화하였다.

앞서 분석된 Group별 변수중요도에서 Group1에서는 도로의 영향력이 가장 크게 나타났는데 도로가 많이 분포(고속도로, 국도 등)되어 있는 외곽지역도 포함하여 독립변수들의 커널밀도에서 또한 전체적으로 퍼져있는 형태로 되어있는 것을 확인하였다. Group2에서 Group4로 갈수록 주거지역과 업종(학문/교육, 부동산, 음식 등), 학령인구, 고령인구의 중요도가 높아지기 시작하였으며 각 독립변수들의 커널밀도함수에서 또한 대구의 중심부인 인구, 주거지, 업종 등의 밀도가 큰 대구의 중심 부분에 밀집되어 있는 것을 확인할 수 있었다. Group별 실거래가의 공간적 분포와 독립변수들의 커널밀도함수는 비슷한 경향을 보여 이와 같은 중요 독립변수들의 커널밀도함수 분포는 향후 실거래가가 공간적으로 어떻게 분포될지 예측할 수 있는 기초자료로 활용할 수 있다.

4. 결 론

부동산은 생활을 영위함에 있어서 반드시 필요한 요인이며, 부동산의 주변 환경에 따라서 가격에 대한 차이가 많이 발생하고 있다. 이러한 부동산의 실거래가를 머신러닝의 한 기법인 XGBoost를 적용하여 부동산 실거래가에 미치는 요인들을 찾아 분석하고자 하였다.

분석에 사용된 데이터는 2020년 7월부터 2023년 6월까지의 대구광역시 국토교통부의 부동산 실거래가, 지목, 용도지역, 행정안전부의 인구통계, 소상공인시장진흥공단의 사업체 수의 통계자료로 설정하였다. 수집된 데이터를 통해 실거래가 통계분포를 확인하기 위해 탐색적 데이터 분석을 수행

15) 국토연구원, “시군구별 노인인구 및 총인구 변화와 시사점(2000~2021년)”, 「국토이슈리포트」, 2022, pp.4-6.

하였으며, 통계분포 결과 ㎡당 200만 원에서 600만 원의 가격이 가장 많이 분포되어 있는 것으로 나타났다. 평균적으로 428.3만 원으로 일반적으로 사용하는 통계적 분위수를 활용하여 Group을 구분한 결과, 1사분위수는 318.87만 원, 2사분위수는 396.02만 원, 3사분위수는 492.31만 원으로 나타났다. 이를 사분위수 별로 Group1(min~1사분위수), Group2(1사분위수~2사분위수), Group3(2사분위수~3사분위수), Group4(3사분위수~max)의 총 4개의 Group으로 설정하여 각각 실거래가의 시공간적 분포에 따라 독립변수의 영향 정도의 차이를 알아본 결과 Group1은 도로, 50~59세의 인구, 0~9세의 인구, 20~29세의 인구, 제3종일반주거지역의 순서, Group2은 제1종일반주거지역, 30~39세, 고령인구, 제3종주거지역 순서로, Group3은 제1종일반주거지역, 부동산업종, 관광/여가/오락 업종, 고령인구 순서로, Group4은 고령인구, 제1종일반주거지역, 10~19세인구, 20~39세 인구 순으로 각각 영향이 큰 요인이 상이한 것으로 나타났다.

각 그룹의 요인분석 결과를 바탕으로 공간적 분포를 확인하기 위해 국토정보플랫폼에서 제공하는 격자별 총인구수 데이터를 활용하여 행정동 인구밀도의 중심점을 산출하였고 그 중심점을 기준으로 독립변수들의 커널밀도를 산출하여 공간적으로 시각화를 진행하였다. Group1에서는 변수중요도의 요인과 유사하게 도로의 영향을 많이 받는 지역에 분포되어 있었고, Group2에서 Group4로 옮겨질수록 변수중요도에서 업종 및 인구의 영향도가 커짐을 확인하였는데 공간적인 변수요인의 밀도가 점점 중심지로 밀집되는 현상이 나타났다. 또한 실거래가의 그룹별 분포와 중요 변수들의 커널밀도함수는 공간적으로 유사하게 나타나 중요 변수들의의 커널밀도함수 분포는 향후 실거래가가 공간적으로 어떠한 특징으로 분포될 수 있는지 예측할 수 있는 기초자료로 활용될 수 있고, 실거래가를 예측하기 위해 중요 변수들을 사용하여 커널밀도함수 또는 다양한 머신러닝, 딥러닝 기술 등을 통해 앞으로의 실거래가 시계열 예측이 보다 더 정확하게 이루어질 수 있다. 향후 연구에서는 요인분석과 함께 적합한 모델 파라미터의 선정을 위해 grid search와 같은 알고리즘을 활용하여 모델을 고도화하는 연구가 필요하다. 또한 예측된 실거래가를 활용하여 부동산 거래가 전략적으로 이루어지거나 정책적으로 필요한 의사결정을 하는데 중요한 자료로 활용될 것으로 예상된다.

〈참고문헌〉

1. 강성경, “증거기반정책을 위한 행정자료의 탐색적 데이터 분석(EDA)과 활용 : 선박안전 분야 관련 속성 분석 및 정책 적용을 중점으로”, 박사학위논문, 동국대학교 대학원, 2021.
2. 고주형·강명구, “부동산 가격 요인과 가격상승률 요인 비교 연구:서울시 재건축 아파트를 중심으로”, 「부동산학연구」, 제25집 2호, 2019.
3. 국토연구원, “시군구별 노인인구 및 총인구 변화와 시사점(2000~2021년)”, 「국토이슈리포트」, 2022.
4. 김학현·유환규·오하영, “딥러닝과 머신러닝을 이용한 아파트 실거래가 예측”, 「정보처리학회 논문지 소프트웨어 및 데이터 공학」, 제12권 2호, 2023.
5. 배성완·유정석, “딥 러닝을 이용한 부동산가격지수 예측”, 「부동산연구」, 제27집 3호, 2017.
6. 국토교통부, https://rt.molit.go.kr/
7. 공공데이터 포털, https://www.data.go.kr/
8. 송기욱·류강민, “헤도닉가격모형을 이용한 물류부동산의 실거래가 결정요인 실증분석”, 「부동산학연구」, 제25집 3호, 2019.
9. 소상공인시장진흥공단, https://www.data.go.kr/data/15083033/fileData.do
10. 행정안전부, https://jumin.mois.go.kr/
11. https://ko.wikipedia.org/wiki/XGBoost
12. M Kalinic, “Kernel Density Estimation (KDE)

vs. Hot-Spot Analysis - Detecting Criminal Hot Spots in the City of San Francisco", 2018.

13. Tukey, "Exploratary Data Analysiss", Journal of the Royal Statistic Society, Vol. 28(1), 1979.

(접수일 2023.11.10., 심사일 2023.11.14., 심사완료일 2023.11.24.)