주택가격 예측을 위한 주요 특성 분석

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

Analysis of Important Features for Predicting House Prices

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

● 요 약 ●

불안정한 부동산 가격은 지속적인 사회 문제로 거론되고 있는데 이는 부동산 매매 가격을 예측할 수 있는 정확한 지표가 체계적이고 구체적으로 확립되지 않았기 때문이다. 본 논문은 가격변동에 주요하게 영향을 미치는 특성을 파악하여 가격 예측 지표로 활용하기 위해 머신러닝 모델을 적용하여 특성 분석을 수행한다. 이를 위해 한국부동산원에서 제공하는 2021년 10월부터 2022년 9월까지 1년간의 역 주변 500M 이내 거래 데이터 약 30만 6천 개를 어떠한 과정으로 전처리하여 머신러닝 모델에 적용하였는지 기술한다.

키워드: 머신러닝(machine learning), 주택가격예측(housing price prediction), 데이터전처리(data preprocessing)

I. Introduction

지속적인 사회문제로 거론되고 있는 불안정한 부동산 특히 주택 가격은 그 원인이 여러 요인에 기인하지만 가장 큰 요인은 가격을 결정하는 특성들의 변동성이 크며 체계적이고 구체적이지 않다는 것이다. 가격 결정의 다양한 특성들 중 정책적인 특성을 제외하고 변동성이 크지 않은 특성들을 선정하여 안정적으로 가격을 예측할 수 있다면, 주택 가격판단 지표로 해당 특성들을 사용할 수 있을 것이다. 본 논문은 이러한 특성들을 분석하여 가격 예측을 위한 머신러닝 모델에 적용하기까지 어떠한 과정으로 데이터 전처리가 수행되었는지 기술한다.

II. Proposed Scheme

1. Data

본 연구를 위해 사용한 데이터는 한국부동산원에서 제공하는 데이터로, GIS를 활용한 역 주변 500M 이내의 실거래가 공개시스템을 통해 수집된 공동주택 거래 정보와 연계된 데이터[1]이다. 기간은 2021년 10월부터 2022년 9월까지로 거래 데이터 약 30만 6천 개를 이용하였다. Table 1은 주택매매 데이터를 구성하는 변수들에 대한 설명이다.

Table 1. Variables of Housing Sale Data

변수명	설명
SIGUNGU_CD	시군구 코드
EMDL_CD	읍면동리 코드
CLL	지번구분(1:일반,2:산)
MNO	지번(본번)
SNO	지번(부번)
ADRES	주소명(법정동)
HUS_TP	공동주택 유형구분 (아파트, 연립, 다세대, 오피스텔)
COMP_NM	단지(건물) 명칭
BLDG_YEAR	건축년도
FLR	층 정보
XUAR	전용면적(m^2)
CTRT_YRMTH	계약년월
CTRT_DAY	계약일
TRANSCT_TYPE	거래유형(매매,전세,월세)
DLNG_AMOUNT	매매금액(만원)
NEAR_SUBW_NM	인접한 지하철역명
NEAR_SUBW_DIST	지하철까지의 직선거리(m)

2. Data Preprocessing

데이터전처리 첫 번째로 인접 지하철역에 대한 주요도를 반영하였다. 기존 선행 연구들은 생활사회기반시설과의 거리를 요인으로 활용하였으나 같은 지하철이라도 이용량이나 중요성이 확연히 달라서 유동인구에 따라 주요 역들이 결정된다고 판단하고 기존 연구[2, 3]와 다르게 가중치를 부여하였다. 티머니에서 제공하는 2022년 10월 교통카드 이용 통계자료 데이터[4]를 사용하였다. 티머니 데이터는 호선, 역ID, 지하철역, 승차승객수, 하차승객수로 구성되는데, 주요 역을 예측하기 위해 사이킷런 Min-Max Scaler를 적용하였다. 또한 우선순위를 부여하기 위해 승·하차승객수를 합산하여 이용객이 많은 순으로 나열하였으며 정규화를 진행하여 값의 범위를 0 ~ 1로 변환하였다. 만약 1년간의 거래 데이터가 없어 부동산 데이터에서 빠진 지하철 역이 존재한다면 모두 확인하여 우선순위를 부여하기 위해 수도권 지하철 역사정보[5, 6]를 추가하여 이용하였다.

두 번째는 아파트 브랜드 유무에 따른 가격 차이가 있다고 판단하여 브랜드변수를 추가하고자 한국기업 평판연구소에서 제공하는 데이터인 아파트 브랜드 선호도 TOP20을 사용하여 상위 46 개 브랜드 데이터를 반영하였다. 브랜드 이름 중 일부가 포함되어 있으면 Value 1, 포함돼있지 않다면 Value 0을 주어 식별할 수 있도록 하였다.

부동산 가격에 따른 거래량의 경우 Fig. 1.에서 보듯이 데이터 간 편차가 매우 크기 때문에 예측 정확도를 높이기 위해 세 번째 전처리 과정인 로그 스케일링으로 정규화를 수행한다. 정규화가 충족됨을 Fig. 2.의 그래프로 알 수 있다. 네 번째 과정은 공동주택 유형구분 특성에 대한 것으로 제공 데이터에는 아파트, 오피스텔, 연립, 다세대 값 중 하나의 문자값이 저장되어 있다. 해당 변수를 각 값에 대한 독립 칼럼으로 나누어 원-핫 인코딩 방식으로 세분화하였다. 이는 각 주택유형이 가격 예측에 어느 정도의 영향을 미치는지 파악하기 위해서이다.

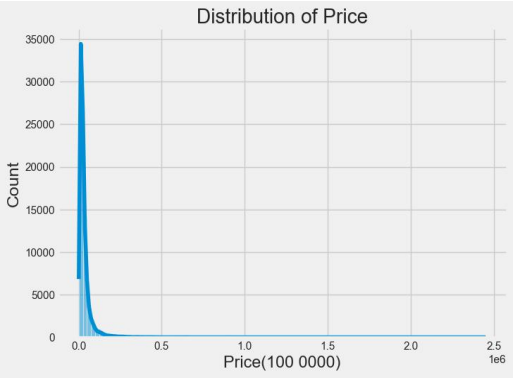

Fig. 1. Trading Volume by Price

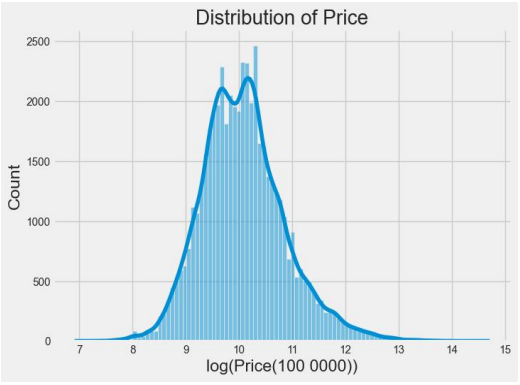

Fig. 2. Log-Normalized Trading Volume by Price

다섯 번째는 건축년도에 대한 전처리 필요이다. 건물연식은 집값 예측에 중요 역할을 한다. 따라서 결측값이 없어야 정확한 예측이 가능하다. 이를 해결하기 사이킷런 KNN impuer 패키지를 사용하여 계산하였다.

3. Mulitcollinearity Test

학습을 수행하기 전 python의 statsmodel 패키지를사용하여 다중공선성(multicollinearity)을 점검하였다. Fig.3에서 보이듯이, 설명변수의 모든 분산팽창지수(vif: variance inflaction factor) 값을 엄격히 적용하여 5 미만으로 기준을 잡아도 모든 변수들이 다중공선성이 없는 것을 확인하였다.

	VIF_Factor	Feature
0	2.610155	FLR
1	4.565436	XUAR
2	4.167406	NEAR_SUBW_DIST
3	3.042181	ele_school
4	1.906565	mid_school
5	2.740688	park
6	1.750695	lib
7	4.570116	지역의따른소득
8	1.229649	COMP_NM_NUM
9	1.985323	지하철역(이용객높을수록큰값)
10	3.056208	age_building
11	7.341679	base_rate

Fig. 3. VIF

4. Modeling Results

머신러닝 모델링 평가지표로 RMSLE를 사용하여, 총 8개의 모델에 전처리한 데이터를 적용하여 비교하였다. Fig. 5에 여러 예측 모델 중 가장 높은 정확도를 보인 모델은 LGBMRegressor 모델로 0.248의 rmsle 값을 갖는다. 이는 낮은 정확도인 0.502를 갖는 LinearRegression 모델보다 2배 이상의 정확도이다. 8개 모델 중 전체적으로 경사하강법 기반의 모델들이 좋은 성능을 보인다. LGBMRegressor 모델에 적용한 데이터세트의 특성 중요도를 분석했을 때, 500M 이내의 역세권 데이터인 것을 감안하면 부동산 가격 결정의 가장 큰 특성은 지하철역이 우선시 되었으며, 건물의 노후 정도, 면적 순으로 나타났다. Fig. 4는 중요 특성 순으로 나열된 그래프이다.

Fig. 4. Features Contribution.

III. Conclusions

본 논문은 한국 부동산원의 역 500M 이내 거래 데이터를 이용하여 주택매매 가격에 영향을 미치는 특성을 찾아보는 연구를 진행하였다. 현재 주택 거래 데이터의 주소를 이용하여 주변의 학교, 공원, 도서관 같은 교육 및 복지시설의 유무와 한국은행 기준 금리를 이용한 특징들을 추가하여 가격 예측의 정확도를 상승하는 여러 연구들이 진행되고 있다. 더불어 가격 변동 특성 중 정책적 및 사회적 특성까지 추가하여 예측할 수 있게 된다면 주택매매 시 정확한 판단 지표로 활용이 가능할 것으로 보인다. 본 연구는 추후 해당 방법들을 적용한 매물 추천 서비스 구축으로 확장하고자 한다.

ACKNOWLEDGEMENT

이 논문은 2021년도 정부 (과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 이공분야 기초연구사업임 (No. 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 Bea. Forecasting Property Prices Using the Machine Learning Methods: Model Comparisons. Dissertation, Dankook University Graduate School, 2019.

[3] Seung-Ju Beak. Predicting the Housing Purchase Price Index using Deep Learning Model. Dissertation, Sungkyunkwan University, 2021.

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

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

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