Asia Pacific Academy of Science Pte. Ltd.

City Diversity 2024, 5(1), 2742.
https://doi.org/10.54517/cd.v5i1.2742

Article

Prediction of new housing prices in Changsha urban area based on multiple machine learning algorithms: A comparative analysis

Junjia Yin[1,2,*], Aidi Hizami Alias[1,2], Nuzul Azam Haron[1], Nabilah Abu Bakar[1]

[1] Department of Civil Engineering, Faculty of Engineering, Universiti Putra Malaysia, Serdang 43300, Malaysia
[2] UPM-Bentley BIM Advancement Lab, Universiti Putra Malaysia, Serdang 43300, Malaysia
*** Corresponding author:** Yin Junjia, gs64764@student.upm.edu.my

CITATION

Yin J, Alias AH, Haron NA, Bakar NA. Prediction of new housing prices in Changsha urban area based on multiple machine learning algorithms: A comparative analysis. City Diversity. 2024; 5(1): 2742. https://doi.org/10.54517/cd.v5i1.2742

ARTICLE INFO

Received: 3 June 2024
Accepted: 11 July 2024
Available online: 5 August 2024

Abstract: As China's pillar industry, the property market has suffered a considerable impact in recent years, with a decline in turnover and many developers at risk of bankruptcy. As one of the most concerned factors for stakeholders, housing prices need to be predicted more objectively and accurately to minimize decision-making errors by developers and consumers. Many prediction models in recent years have been unfriendly to consumers due to technical difficulties, high data demand, and varying factors affecting house prices in different regions. A uniform model across the country cannot capture local differences accurately, so this study compares and analyses the fitting effects of multiple machine learning models using February 2024 new building data in Changsha as an example, aiming to provide consumers with a simple and practical reference for prediction methods. The modeling exploration applies several regression techniques based on machine learning algorithms, such as Stepwise regression, Robust regression, Lasso regression, Ridge regression, Ordinary Least Squares (OLS) regression, Extreme Gradient Boosted regression (XGBoost), and Random Forest (RF) regression. These algorithms are used to construct forecasting models, and the best-performing model is selected by conducting a comparative analysis of the forecasting errors obtained between these models. The research found that machine learning is a practical approach to property price prediction, with least squares regression and Lasso regression providing relatively more convincing results.

Keywords: property market; lasso regression; ridge regression; extreme gradient boosted regression; robust regression; house price forecast; random forest; machine learning

1. Introduction

The real estate industry is fundamental in the economic systems of many developing countries such as China and Malaysia [1], directly impacting gross domestic product (GDP) and employment. Through real estate development, urban infrastructure is improved, and residents' quality of life is enhanced. The development of it helps stimulate domestic demand and positively impacts related sectors such as construction materials, stock exchange, furniture, and home appliances. According to the latest data from the National Bureau of Statistics [2], as shown in **Figure 1**, China's housing market has changed dramatically in terms of its share and size of GDP. Nowadays, the main policy direction of "houses are for living, not for speculation" has not changed, and the capital of real estate enterprises is facing a tight situation [3]. Many enterprises, such as Evergrande Real Estate, have begun to be exposed to a massive debt crisis. Such changes are being directly reflected in the evolution of residential prices. High house prices are detrimental to consumer welfare, while low house prices are detrimental to government revenue [4]. Therefore, observing and

predicting house prices has always been a hot topic in economics, given the importance of China's macroeconomy, individual buyers, and the market. Rico-Juan and Taltavull de La Paz [5] found that the results of prediction models based on home sales data during booms and recessions differed significantly. In other words, each different economic cycle requires new forecasting models. It also means that the prediction model built by, e.g., Xu and Zhang [6], based on the monthly house price dataset of 99 major cities in China from June 2010 to May 2019, can no longer reflect the current house price situation.

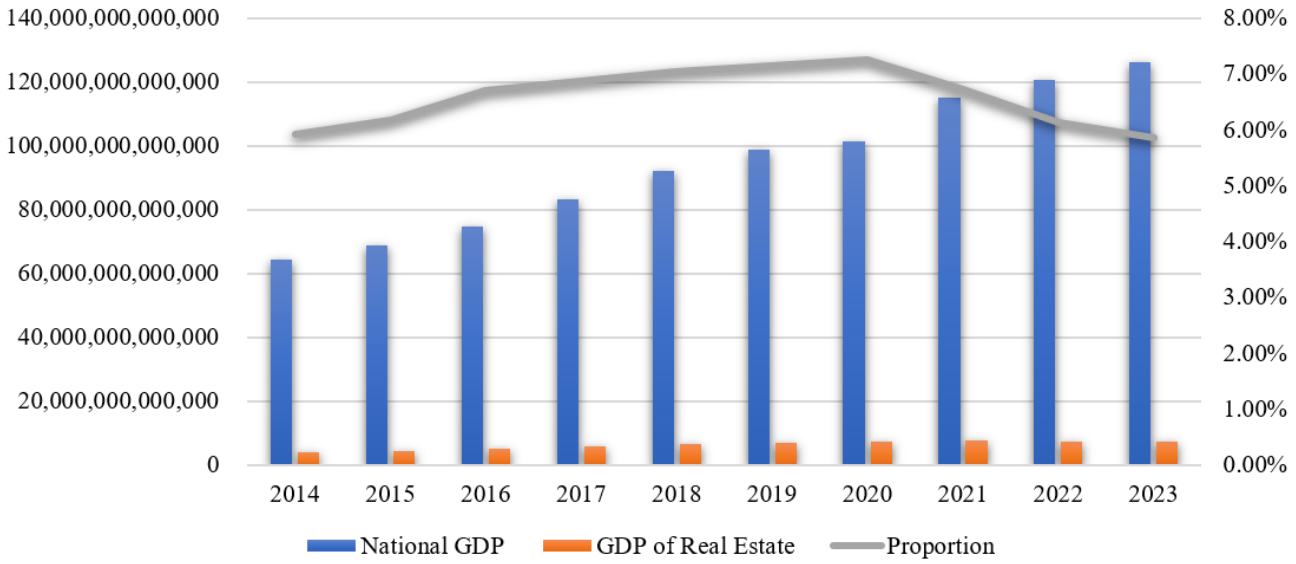

Figure 1. Contribution of real estate to GDP in China, 2014–2023.

Traditionally, there has been much controversy among those who have favored using traditional models and hedonic or repeat sales techniques to predict house prices. One study concluded that only wind speed affects home prices but not the distance from landfills [7]. Miles and Monro's research in the UK identified changes in the risk-free real interest rate as the primary driver of house price changes, with a sustained 1% rise in index-linked gilt yields from current rates ultimately leading to a fall in accurate house prices of around 20% [8]. Many studies are accustomed to analyzing house prices with macro conditions such as credit markets, house price expectations, financial stability, and monetary policy [9]. Barron et al. found that for every 1% increase in listings on US rental apps, rents rose by 0.018% and house prices by 0.026% [10]. Bangura and Lee's study demonstrates that regional differences in house prices in Sydney, for example, can be disregarded without considering macro-factors such as economic policy, lending rates, and overall individual buyers' confidence, as these are already evenly reflected in the raw house price data. In other words, these factors are highly consistent across the same city [11]. In addition, with the development of machine learning, the applicability of some new methods, such as Random Forest and XGBoost algorithms, is equally worthy of attention. Most scholars still construct prediction models based on multiple linear regression. For example, Liu used the least squares method to solve the model's unknown parameters with a maximum error of no more than 8% [12]. Comparative analysis of house price models of machine learning algorithms is an effective way to find the optimal prediction

method [13–15].

Therefore, this study aims to explore and develop a low-cost, easy-to-use forecasting model for regional house prices rather than macro-city average house prices. The broad property market is generally divided into three levels: the land market, the new housing market, and the secondary housing market. In this study, only the new housing market is discussed, as consistency of evidence is beneficial to ensure the study's credibility. The main innovations of this study are threefold: (1) the construction of a readily accessible system of quantitative factors affecting real estate prices; (2) the development of several machine-learning-based price prediction models, and the identification of the most influential and best model through comparative analyses; and (3) the study area is Changsha, a new first-tier city in China, where no similar studies have been conducted, and the sample size covers all the new real estate properties for sale in that month.

The remainder of this study is organized as follows. Section 2 describes the methodology used in this study. Section 3 shows the results of each prediction model with comparative analyses. Section 4 proposes strategies to promote healthy real estate development in China. Section 5 summarizes this study and clarifies the research limitations.

2. Materials and methods

2.1. Collection of sample data

Changsha is the capital of Hunan Province, China; located in the north-eastern part of Hunan province, bordering the lower reaches of the Xiangjiang River and belonging to the western edge of the Xiangliu Basin, it has a subtropical monsoon climate with a mild climate and abundant precipitation [16]. With an area of 11,819 square kilometers, it is also China's 17th most populous city, with a population of over 10 million, and the third most populous city in central China. Geographic features include mountains, hills, and plains [17]. In February, we collected data from 195 new properties for sale through the Anjuke website [18], covering almost all properties for sale in Changsha's urban areas. The specific areas are illustrated in **Figure 2**.

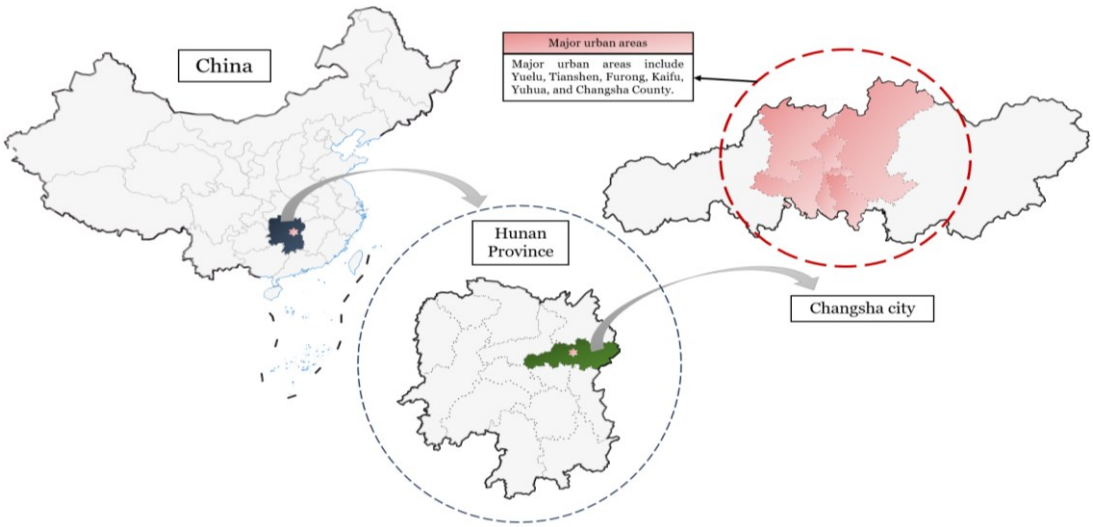

Figure 2. Location map of the study.

Figure 3 illustrates the changes in average house prices in Changsha from 2015 to 2024. The average house price in Changsha was only around 6000 in 2015 but saw explosive growth in 2018, peaking at 11,000. After 2020, it entered a long period of volatility, with prices now receding to around 10,000. This is due primarily to Changsha's relatively stable economic base, which attracts large inflows of people and speculative buyers. These population inflows have led to increased housing demand, pushing housing prices. However, after 2018, the government introduced a series of real estate market control policies, including purchase, sale, and price restrictions. These policies played a role in curbing the rise in housing prices. With the outbreak of the COVID-19 epidemic in 2020, investment enthusiasm waned, some speculative homebuyers began to withdraw, and housing prices declined as a result. Although starting in 2023, the government adopted some supportive policies to stabilize the market and stimulate investment enthusiasm and speculative behavior. The effect is still insignificant.

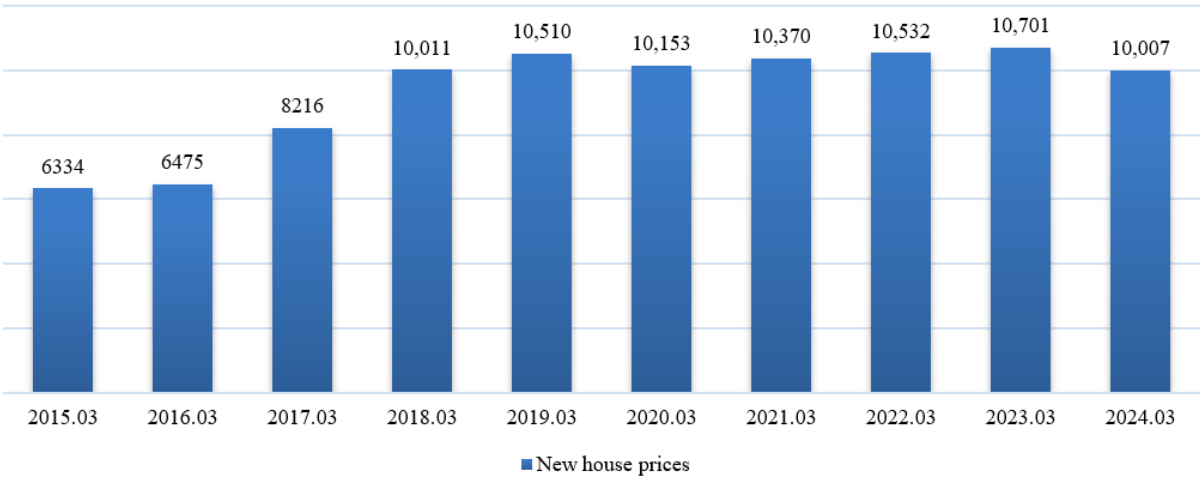

Figure 3. Average house price in Changsha, 2015–2024.

For the same city, factors such as climate, policies, resident confidence, and cultural characteristics remain broadly consistent and are not listed for discussion in this study. The difference between each property lies in the different community product parameters, including plot ratio, green ratio, car parking ratio, neighborhood amenities, etc. This study aims to establish a consumer-friendly quantitative indicator system. In other words, consumers can readily access precise information through relevant home-buying websites. Therefore, some data indicators that are hard to obtain or quantify are not included in this indicator system. Eventually, the following 18 quantitative indicators were selected for this study: site area, floor area, ring road location, developer popularity, degree of renovation, building type, property fee, number of planned households, greening rate, floor area ratio, car parking space, total building, car park ratio, number of house types, number of public transport within 3 km, number of shops within 3 km, number of schools within 3 km, and number of hospitals within 3 km. It is worth noting that the scientific validity regarding the selection of indicators has been verified through relevant literature [19–22], as presented in **Table 1**. The study hypothesizes that all these factors have an impact on house prices.

Table 1. Quantitative data categories.

ID	Items	Explanation
RP	Residential price	The unit price of a house is the price per square meter, usually expressed as "yuan/m^2".
LA	Land area	The building footprint is the area a building or structure occupies in a horizontal plane. It is usually measured in square meters (m^2).
BA	Building area	The building area, also known as the building unfolded area, is the sum of the plan areas of each residential building floor measured at the level of the perimeter above the footings of the external walls of the building.
LI	Loop location	A ring road is a circular transport system made up of roads, which is a sign of whether the area where a house is located is remote. Changsha has three ring roads, named 1, 2, and 3, from near to far.
DV	Developer visibility	Developer awareness refers to the degree to which a property developer is known in the marketplace and how the public perceives its brand. It is categorized as well-known (1) or common (0).
DE	Decoration	Renovation is carried out in a specific area and scope, including plumbing and electrical work, walls, floors, ceilings, landscaping, etc. This study is divided into unfurnished (0) and furnished (1).
BC	Building category	The structural categories of buildings discussed in this study are towers (1), slabs (2), and slab-tower combinations (3), which are common structural categories of residential buildings.
PT	Property fee	The property fee is the fee the property manager charges the owner or occupier for the services provided under the property service contract, usually expressed as "yuan/m^2".
PH	Planned households	The number of people planned to live in the neighborhood.
GR	Green ratio	The greening ratio is the ratio of the vertical projection area of green plants to the area of green space.
PR	Plot ratio	Floor Area Ratio (Gross Floor Area Density) is the ratio of a neighborhood's total above-ground floor area to the net site area.
CP	Car park	Number of car parking spaces.
TB	Total buildings	Total number of buildings.
CR	Car park ratio	The parking ratio is between the "total number of households" and the "total number of parking spaces" in a neighborhood.
NH	Number of house types	The number of house types refers to the number of different house types in a property project. Developers usually plan various house types to meet the needs of other home buyers.
ME	Number of public transport within 3 km	It refers to the number of public transport stops or routes within 3 km of a location. This may include bus stops, metro stations, tram stops, etc.
TR	Number of shops within 3 km	It indicates the number of shops, supermarkets, convenience stores, etc., within 3 km of a given location.
SC	Number of schools within 3 km	It is the number of schools of all types, including primary, secondary, and tertiary, within 3 km of a given location.
HO	Number of hospitals within 3 km	It is the number of hospitals, clinics, health posts, etc., within 3 km of a given location.

The detailed descriptive analyses of the 195 cases are shown in **Table 2**. In terms of house price (RP), for example, the current fluctuation of house prices for each new development is significant, ranging from 5000 to 29,000.

Table 2. Descriptive statistics.

Items	*N*	Minimum	Maximum	Mean	Std. deviation	Variance	Skewness		Kurtosis	
							Statistics	Standard error	Statistics	Standard error
RP	195	5000	29,000	12,216.90	4095.975	16,777,009.333	1.311	0.172	3.083	0.343
LA	195	3923.00	1,100,000.00	104,390.496	132,756.323	17,624,241,549.996	3.824	0.172	19.875	0.343
BA	195	8736.00	2,034,133.10	299,652.368	303,566.887	92,152,855,206.518	2.962	0.172	11.077	0.343
LI	195	1	3	1.99	0.537	0.288	−0.005	0.172	0.534	0.343

Table 2. (*Continued*).

Items	*N*	Minimum	Maximum	Mean	Std. Deviation	Variance	Skewness		Kurtosis	
							Statistics	Standard error	Statistics	Standard error
DV	195	0	1	0.56	0.498	0.248	−0.234	0.172	−1.965	0.343
DE	195	0	1	0.34	0.475	0.226	0.673	0.172	−1.563	0.343
BC	195	1	3	1.78	0.915	0.837	0.443	0.172	−1.667	0.343
PT	195	1.60	4.60	2.802	0.566	0.321	0.470	0.172	0.355	0.343
PH	195	59	6933	1395.93	1191.169	1,418,883.566	1.979	0.172	5.131	0.343
GR	195	25.00%	64.00%	37.35%	4.83%	23.372	0.668	0.172	4.684	0.343
PR	195	1.14	7.34	2.923	1.035	1.071	1.289	0.174	2.503	0.346
CP	195	70	8914	1562.76	1368.120	1,871,753.352	2.520	0.173	9.100	0.344
TB	195	1	197	18.86	22.096	488.239	4.948	0.172	34.011	0.343
CR	195	0.244	7.988	0.982	0.627	0.393	7.522	0.173	79.815	0.344
NH	195	1	92	7.30	8.834	78.048	7.293	0.172	63.708	0.343
ME	195	0	28	6.81	5.252	27.580	1.311	0.172	2.383	0.343
TR	195	0	25	4.89	4.676	21.867	1.578	0.172	3.235	0.343
SC	195	6	100	70.68	30.603	936.543	−0.599	0.172	−1.115	0.343
HO	195	0	71	10.07	13.469	181.409	2.476	0.172	6.493	0.343

2.2. Regression analysis

2.2.1. Stepwise regression

Stepwise regression is a statistical method that builds regression models by automatically selecting predictive variables [23]. Here are the main approaches for stepwise regression [24]:

1) Forward selection: Start with no variables in the model. Test the addition of each variable using a chosen model fit criterion. Add the variable (if any) that gives the most statistically significant improvement to the fit. Repeat until no further improvements occur.
2) Backward elimination: Begin with all candidate variables. Test the deletion of each variable using a chosen model fit criterion. Remove the variable (if any) that results in the least statistically significant deterioration of the model fit. Repeat until no more variables can be deleted without substantially losing fit.
3) Bidirectional elimination: Combines forward selection and backward elimination. At each step, test variables for inclusion or exclusion.

Stepwise regression helps manage the complexity of model building, especially when dealing with many potential explanatory variables.

2.2.2. Ridge regression

Ridge Regression is a method for estimating the coefficients of a multiple regression model for situations where the independent variables are highly correlated with each other [25]. The theory of ridge regression was first proposed by Hoerl and Kennard in 1970, resulting from a decade of research in ridge analysis. It addresses the inaccuracy of least squares estimators in linear regression models with multicollinear (highly correlated) independent variables. Ridge regression provides more accurate estimates of the ridge parameters by creating a ridge regression

estimator (RR), whose variance and mean-square estimates are typically smaller than those of the previous least-squares estimator but which introduces a certain amount of bias (see bias-variance trade-off). It is used in many fields, including econometrics, chemistry, and engineering.

Ridge Regression has two steps [26]. Step 1: Before ridge regression analysis, the *K*-value needs to be confirmed in conjunction with the ridge trace plot; the *K*-value is chosen as the smallest *K*-value at which the standardized regression coefficients of the independent variables converge; the smaller the *K*-value, the smaller the bias. When the *K*-value of 0 is an ordinary linear OLS regression. The smaller the *K*-value, the better it is, and it is usually recommended to be less than 1; after determining the *K*-value, the *K*-value can be actively input to produce the ridge regression model estimation.

2.2.3. Robust regression

Robust Regression (RR) is a method used to fit regression models designed to overcome some of the limitations of traditional regression analysis [27]. Standard regression methods (e.g., ordinary least squares) have good properties when their underlying assumptions are valid. Still, the results can be misleading if the assumptions are invalid (i.e., not robust to assumption violations). It aims to mitigate the impact of assumption violations on regression estimates during data generation.

One common problem is heteroscedastic errors (Heteroscedastic errors). In a homoscedastic model, the variance of the error term is assumed to be constant for all values of *x*. In a heteroscedastic model, the variance of the error term is assumed to be constant for all values of *x*. Heteroscedastic models, on the other hand, allow the variance of the error term to depend on x, which is more consistent with many practical situations. Another common situation is the presence of outliers. Robust estimation methods can be applied when the data contains outliers [28]. Ordinary least squares are very sensitive to the estimation of the regression model, and the magnitude of error for outliers is twice that of typical observations, thus contributing four times as much to the loss of squared error and having a more significant impact on the regression estimate.

The Huber loss function is a robust alternative to the standard squared error loss that reduces the contribution of outliers to the squared error loss, thus limiting its impact on the regression estimates [29].

2.2.4. Lasso regression

Lasso Regression (Least Absolute Value Shrinkage and Selection Operator Algorithm) is a method for regression analysis designed to perform both variable selection and regularization to improve the predictive accuracy and interpretability of the generated statistical model [30]. It also enhances standard linear regression by introducing an L1 norm with a penalty coefficient, λ, as a regularization term. The goal is to minimize the cost function [31]:

$$\text{Cost}\ (w) = \sum_{i=1}^{N} (y_i - w^T x_i)^2 + \lambda \parallel w \parallel_1 \quad (1)$$

where, y_i is the target label of the observation, x_i is the corresponding feature vector, and w is the vector of weight coefficients.

Due to the non-smoothness of the L1 norm, it is not possible to obtain an analytical solution directly by derivation. Therefore, two commonly used solution methods exist. (1) Coordinate descent method: iteratively update the weight coefficients along the direction of each coordinate axis to approximate the optimal solution. (2) Minimum angle regression method: calculate the correlation step by step by feature selection to reduce the number of iterations. The former is adopted in this study.

2.2.5. OLS regression

Ordinary Least Squares (OLS) regression is an optimization strategy for linear regression models that aims to find a straight line as close as possible to the data points. It is a class of algorithms for supervised machine-learning tasks [32]. Unlike classification tasks, the target variables of regression tasks are continuous values, not categories. We predict house prices based on some features, where the output is a constant value. Simple linear regression is a commonly used statistical model for predicting a target variable based on a single feature. It is based on the following formula [33]:

$$Y = \beta_0 + \beta_1 X + \epsilon_i \tag{2}$$

where, Y is the target variable; X is a single characteristic; β_0 and β_1 are regression coefficients and ϵ_i is the error term.

OLS regression estimates regression coefficients by minimizing the squared error between the model and the data points. Specifically, it computes the residual sum of squares (RSS), the sum of the squares of all error values. The estimated value of the regression coefficient minimizes this sum of squares. Typically, OLS regression performs well under the following conditions: the independent variable is exogenous, the residuals have finite variance, and the residuals follow a normal distribution with zero mean.

2.2.6. XGBoost regression

XGBoost (eXtreme Gradient Boosting) is a machine-learning algorithm based on Gradient Boosting Trees [34]. It is popular in Kaggle competitions and real-world applications because of its excellent performance in handling large-scale datasets and complex models. With parallel processing and efficient memory usage, XGBoost can process large-scale data quickly and supports regression, classification, sorting, and custom objective functions. It also prevents overfitting through L1 (Lasso) and L2 (Ridge) regularization.

It is based on gradient boosting, which means that the model is progressively improved by iteratively constructing a tree. The goal of each iteration is to minimize some loss of function. The model formula is as follows:

Suppose we have a training dataset $D = \{(x_i, y_i)\}$, where x_i is the input features and y_i is the corresponding label. We denote the predicted value at the t iteration by $\hat{y}_i^{(t)}$. The prediction function of the model is [35]:

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) \tag{3}$$

where A is the tree constructed at the *k*-th iteration. XGBoost trains the model by minimizing the following objective function:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{4}$$

where l is the loss function (e.g., mean square error, logistic loss, etc.). $\Omega(f_t)$ is the regularization term which is used to control the complexity of the model and is usually defined as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{5}$$

where T is the number of leaf nodes of the tree, w_j is the weights of the leaf nodes and γ and λ are hyperparameters that regulate the complexity of the model.

In each iteration, XGBoost constructs a new tree f_t using a greedy algorithm. This process consists of the following steps:

Calculate first-order and second-order derivatives: compute the first-order derivative B and the second-order derivative $g_i$for the loss function h_i:

$$g_i = \frac{\partial l\left(y_i, \hat{y}_i^{(t-1)}\right)}{\partial \hat{y}_i^{(t-1)}} \tag{6}$$

$$h_i = \frac{\partial^2 l\left(y_i, \hat{y}_i^{(t-1)}\right)}{\partial \hat{y}_i^{(t-1)2}} \tag{7}$$

Selecting the best splitting point: At each node, select the splitting point that causes the objective function to decrease the most. The gain of the objective function is calculated as:

$$\text{Gain} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{(\sum_{i\in\in} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma \tag{8}$$

where A and B denote the set of samples from the left and right child nodes, respectively.

Update leaf node weights: for each leaf node, calculate the optimal weights C.

$$w_j = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda} \tag{9}$$

By iterating the above process, it can gradually improve the model's prediction accuracy and achieve better results.

2.2.7. Random forest regression

Random Forest is an integrated learning algorithm that improves the accuracy and stability of a model by constructing multiple decision trees and combining their predictions [36]. It performs well in classification and regression tasks and has the advantages of handling high-dimensional data, reducing overfitting, and being computationally efficient.

The basic idea of a Random Forest is to construct a set of independent decision trees by introducing randomness and using the collective intelligence of these trees to make predictions. Random Forest contains the following two critical steps in the construction process:

1) Bootstrap sampling: multiple subsample sets are drawn from the original training set with a putback, each used to train a decision tree.

2) Feature Random Selection: during the splitting process of each node, a portion of the features are randomly selected to determine the optimal splitting point instead of using all the features.

These two steps increase the diversity of the model and reduce the correlation between the trees, thus improving the generalization of the overall model. Random forest construction process

Bootstrap sampling: assuming the original training set is $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, from which we have put back to draw B sample sets, each of which is of size n. Training decision trees: for each sample set, a decision tree is constructed. In constructing each tree, feature random selection is performed. Prediction: for new input data, the final output is obtained by integrating the prediction results of each decision tree. In classification tasks, majority voting is used; the average is taken in regression tasks.

In the regression task, it is assumed that the prediction for each tree is $\hat{y}_i^{(b)}$, and the final prediction $\hat{y}_i$ is the average of the predictions for each tree [37]:

$$\hat{y}_i = \frac{1}{B}\sum_{b=1}^{B} \hat{y}_i^{(b)} \tag{10}$$

3. Results

Table 3 shows the results of the Pearson correlation analysis. This study visualizes the relationship between the influencing factors through **Figure 4**. Due to space constraints, this study has only analyzed the relationship between house prices and other factors in depth. According to the discriminant criteria of the study [38], the results of the Pearson correlation analysis were significant. Hence, the research assumptions were valid, and it was possible to proceed to the next step of the regression analysis.

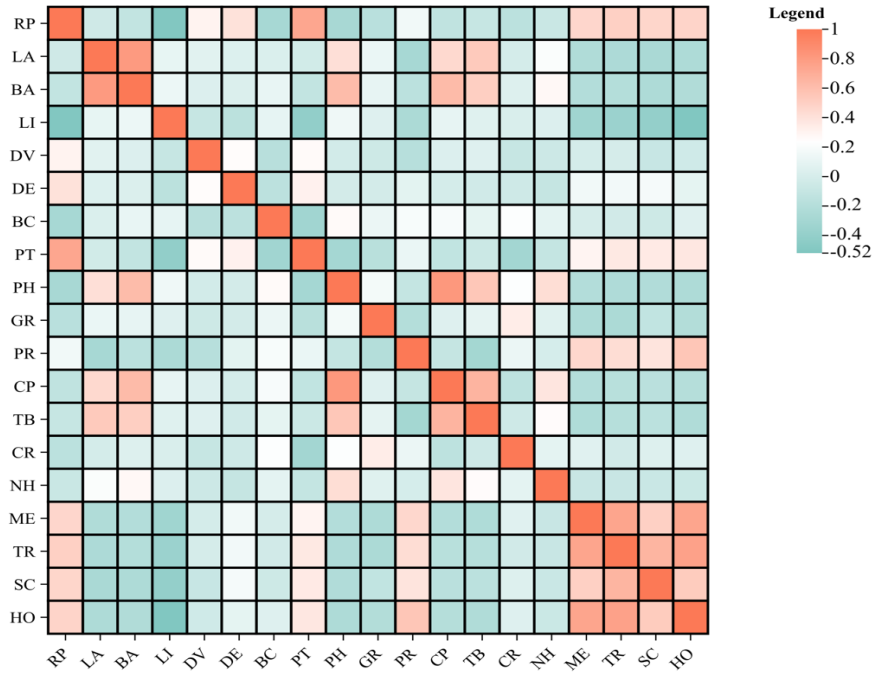

Figure 4. The heatmap of Pearson correlation analysis.

Table 3. Pearson correlation analysis results.

7		LA	BA	LI	DV	DE	BC	PT	PH	GR	PR	CP	TB	CR	NH	ME	TR	SC	HO
RP	PC	−0.052	−0.128	−0.515**	0.309**	0.405**	−0.288**	0.742**	−0.284**	−0.182*	0.155*	−0.139	−0.104*	−0.168*	−0.085*	0.470**	0.505**	0.474**	0.485**
	ST	0.065	0.072	0.000	0.000	0.000	0.000	0.000	0.000	0.010	0.030	0.051	0.044	0.018	0.032	0.000	0.000	0.000	0.000

Note: **. The correlation is significant at the 0.01 level (two-tailed); *. Significant at the 0.05 level (two-tailed); Pearson correlation (PC); Significance two-tailed (ST).

3.1. Results of stepwise regression

The difference between stepwise regression and regression analysis is that a stepwise regression model automatically identifies significant independent variables (*X*), and *X* that is not significant is automatically moved out of the model. As reflected in **Table 4**, the first step is to analyze the model fit R^2, as well as the VIF value (or tolerance value, tolerance = 1/VIF value) can be analyzed (to determine the multiple covariances, VIF > 5 generally indicates that there is covariance or tolerance < 0.2 shows typically that there is a covariance problem) [37]. Step 2: Write the model equation. Step 3: Analyze the significance of *X*. If it is significant, *X* has an influential relationship with *Y*.

Table 4. Results of stepwise regression analysis (n = 195).

Items	Non-standardized coefficients		Standardized coefficient	*t*	*p*	Covariance diagnostics	
	B	**Standard error**	**Beta**			**VIF**	**Tolerability**
Constant	1221.494	1424.276	-	0.858	0.392	-	-
LI	−1226.439	359.582	−0.160	−3.411	0.001**	1.329	0.752
DV	1140.802	361.957	0.140	3.152	0.002**	1.180	0.847
DE	1067.193	379.410	0.125	2.813	0.005**	1.184	0.845
PT	3653.414	358.618	0.498	10.187	0.000**	1.438	0.696
ME	147.564	37.185	0.190	3.968	0.000**	1.375	0.727
SC	16.965	6.776	0.128	2.504	0.013*	1.562	0.640
R^2	0.687						
Adjusted R^2	0.677						
F	$F(6188) = 68.849$, $p = 0.000$						
D-W value	2.031						

Implicit variable: RP; *$p < 0.05$ **$p < 0.01$.

As independent variables, LA, BA, LI, DV, DE, BC, PT, PH, GR, PR, CP, TB, CR, NH, ME, TR, SC, and HO were used. In contrast, RP was used as the dependent variable for the stepwise regression analysis. After the model was automatically recognized, the final residuals were LI, DV, DE, PT, ME, and SC. A total of 6 items are in the model, and the R^2 value is 0.687, which means that LI, DV, DE, PT, ME, and SC can explain 68.7% of the reasons for the changes in RP. And the model passed the *F*-test ($F = 68.849$, $p = 0.000 < 0.05$), which means the model is valid [39].

In addition, the test for the model's multiple covariance found that all the VIF values in the model are less than 5, which means there is no DV, DE, PT, ME, or SC have a significant positive effect on RP [40]. In addition, LI will have a significant negative influence on RP.

3.2. Results of ridge regression

This study, combined with the ridge trace plot, confirms a *K*-value of 0.8. The ANOVA test (the *F*-test) determines if the model is meaningful. As shown in **Table 5**, a *p*-value (sig value) less than 0.05 means that the model is meaningful.

Table 5. ANOVA analysis (n = 195).

Items	Square sum	df	Mean square	*F*	*p*-value
Regression	2,090,481,122.927	18	116,137,840.163	18.171	0.000
Residuals	1,124,898,373.719	176	6,391,468.032		
Total	3,215,379,496.646	194			

Table 6 shows that the model R^2 value is 0.650, which means that LA, BA, LI, DV, DE, BC, PT, PH, GR, PR, CP, TB, CR, NH, ME, TR, SC, HO explains 65.02% of the variation in RP. The *F*-test of the model found that the model passes the *F*-test (F = 18.171, p = 0.000 < 0.05), which means that it shows that at least one of LA, BA, LI, DV, DE, BC, PT, PH, GR, PR, CP, TB, CR, NH, ME, TR, SC, HO will affect RP.

Table 6. Results of ridge regression (n = 195).

Items	Non-standardized coefficients		Standardized coefficient	*t*	*p*	VIF
	B	Standard error	Beta			
Constant	7645.450	1135.511	-	6.733	0.000**	-
LA	0.001	0.001	0.037	1.735	0.084	0.226
BA	0.000	0.000	0.001	0.074	0.941	0.206
LI	−976.563	182.290	−0.128	−5.357	0.000**	0.286
DV	853.888	198.370	0.105	4.305	0.000**	0.297
DE	986.092	207.930	0.115	4.742	0.000**	0.297
BC	−301.640	107.625	−0.068	−2.803	0.006**	0.293
PT	1873.633	169.826	0.256	11.033	0.000**	0.270
PH	−0.185	0.066	−0.055	−2.822	0.005**	0.188
GR	−1451.218	2061.558	−0.017	−0.704	0.482	0.295
PR	−79.419	91.126	−0.020	−0.872	0.385	0.271
CP	0.000	0.055	0.000	0.004	0.997	0.171
TB	0.317	4.131	0.002	0.077	0.939	0.257
CR	−150.983	150.851	−0.023	−1.001	0.318	0.273
NH	5.526	11.139	0.012	0.496	0.620	0.299
ME	71.973	16.822	0.093	4.278	0.000**	0.236
TR	72.077	17.880	0.082	4.031	0.000**	0.210
SC	13.037	3.066	0.098	4.252	0.000**	0.268
HO	23.550	6.110	0.077	3.854	0.000**	0.200
R^2	0.650					
Adjusted R^2	0.614					
F	$F(18,176) = 18.171$, $p = 0.000$					

Implicit variable: RP; *p < 0.05 **p < 0.01.

3.3. Results of robust regression

Robust regression will give different weights to different points; if the residual value of a point is small, it will be given a higher weight, and the residual value of the abnormal point will be more significant because its corresponding weight will be small. The final fit is also more robust and reliable by giving different weights to the residuals of various points. Many scholars have proposed their strategies for constructing the weights, and this study uses the more common Huber's method (the default *t*-value is taken as 1.345) for robust regression.

As reflected in **Table 7**, Robust regression found that DV, DE, PT, CP, CR, and ME would have a significant positive relationship on RP, and LI, PH, and PR would have a substantial adverse effect relationship on RP. However, LA, BA, BC, GR, TB, NH, TR, SC, and HO will not affect RP.

Table 7. Results of robust regression (n = 195).

Items	Standardized coefficient	Standard error	*t*	*p*	95% CI	R^2	Adjusted R^2	*F*
Constant	4862.410	1771.504	2.745	0.006**	1390.327–8334.493			
LA	0.003	0.002	1.837	0.066	−0.000–0.007			
BA	−0.001	0.001	−0.605	0.546	−0.002–0.001			
LI	−1126.517	298.791	−3.770	0.000**	−1712.137–−540.897			
DV	693.139	281.095	2.466	0.014*	142.203–1244.074			
DE	1073.665	294.081	3.651	0.000**	497.276–1650.054			
BC	−131.520	158.787	−0.828	0.408	−442.738–179.698			
PT	3102.193	308.149	10.067	0.000**	2498.233–3706.153			
PH	−0.712	0.263	−2.707	0.007**	−1.227–−0.196			
GR	−3485.930	2996.968	−1.163	0.245	−9359.879–2388.019	0.700	0.670	$F(18,176) = 22.837$ $p = 0.000$
PR	−435.982	171.035	−2.549	0.011*	−771.205– −100.759			
CP	0.563	0.254	2.216	0.027*	0.065–1.061			
TB	−14.009	8.820	−1.588	0.112	−31.296–3.278			
CR	702.583	300.167	2.341	0.019*	114.267–1290.899			
NH	18.215	15.783	1.154	0.248	−12.720–49.149			
ME	134.934	40.663	3.318	0.001**	55.237–214.632			
TR	43.096	52.920	0.814	0.415	−60.625–146.818			
SC	11.415	5.976	1.910	0.056	−0.297–23.127			
HO	19.468	19.162	1.016	0.310	−18.089–57.025			

Implicit variable: RP; *$p < 0.05$ **$p < 0.01$.

3.4. Results of lasso regression

The *K*-value is taken as 0.8, and from **Table 8** above, it is observed that the model R^2 value is 0.711, which means that LA, BA, LI, DV, DE, BC, PT, PH, GR, PR, CP, TB, CR, NH, ME, TR, SC, HO explains 71.11% of the variation in RP. When the model was subjected to the *F*-test, it was found that it passed the *F*-test ($F = 24.066$, $p = 0.000 < 0.05$).

Table 8. Results of Lasso regression (n = 195).

Items	Non-standardized coefficients		Standardized coefficient	*t*	*p*
	B	**standard error**	**Beta**		
Constant	3593.099	1031.880	-	3.482	0.001**
LA	0.002	0.001	0.037	2.891	0.004**
BA	0.000	0.000	0.001	0.000	1.000
LI	−1247.133	165.654	−0.128	−7.529	0.000**
DV	951.411	180.266	0.105	5.278	0.000**
DE	1156.869	188.953	0.115	6.123	0.000**
BC	−119.450	97.803	−0.068	−1.221	0.224
PT	3442.159	154.327	0.256	22.304	0.000**
PH	−0.738	0.060	−0.055	−12.361	0.000**
GR	−2371.863	1873.412	−0.017	−1.266	0.207
PR	−426.712	82.809	−0.020	−5.153	0.000**
CP	0.601	0.050	0.000	11.992	0.000**
TB	−10.327	3.754	0.002	−2.751	0.007**
CR	550.498	137.084	−0.023	4.016	0.000**
NH	20.145	10.122	0.012	1.990	0.048*
ME	131.532	15.287	0.093	8.604	0.000**
TR	56.026	16.248	0.082	3.448	0.001**
SC	14.756	2.786	0.098	5.296	0.000**
HO	10.708	5.553	0.077	1.929	0.055
R^2	0.711				
Adjusted R^2	0.682				
F	$F(18,176) = 24.066$, $p = 0.000$				

Implicit variable: RP; *$p < 0.05$ **$p < 0.01$.

3.5. Results of OLS regression

From **Table 9**, the R^2 value of the model is 0.712, which means that LA, BA, LI, DV, DE, BC, PT, PH, GR, PR, CP, TB, CR, NH, ME, TR, SC, HO, explains 71.17% of the variation in RP. The F-test of the model was found to pass the F-test ($F = 24.204$, $p = 0.000 < 0.05$). It was found that DV, DE, PT, CP, CR, ME, and SC would significantly affect RP. LI, PH, PR, and TB will also significantly negatively influence RP. However, LA, BA, BC, GR, NH, TR, and HO do not affect RP.

Table 9. Results of OLS regression (n = 195).

Items	Standardized coefficient	Standard error	*t*	*p*	95% CI
Constant	3820.653	2224.497	1.718	0.086	−539.282–8180.587
LA	0.002	0.002	1.271	0.204	−0.001–0.006
BA	−0.000	0.001	−0.214	0.831	−0.003–0.002
LI	−1262.340	373.091	−3.383	0.001**	−1993.584–−531.095
DV	946.747	315.260	3.003	0.003**	328.849–1564.644

Table 9. (*Continued*).

Items	Standardized coefficient	Standard error	*t*	*p*	95% CI
DE	1178.706	365.372	3.226	0.001**	462.590–1894.823
BC	−119.971	215.260	−0.557	0.577	−541.873–301.931
PT	3455.932	467.936	7.385	0.000**	2538.795–4373.069
PH	−0.865	0.286	−3.025	0.002**	−1.426–−0.305
GR	−3032.949	4474.595	−0.678	0.498	−11,802.994–5737.095
PR	−481.740	217.967	−2.210	0.027*	−908.948–54.532
CP	0.766	0.253	3.029	0.002**	0.271–1.262
TB	−14.080	6.999	−2.012	0.044*	−27.798–−0.362
CR	704.278	262.516	2.683	0.007**	189.756–1218.801
NH	21.986	15.622	1.407	0.159	−8.632–52.604
ME	132.512	51.391	2.579	0.010**	31.788–233.236
TR	58.561	64.948	0.902	0.367	−68.734–185.857
SC	14.668	6.462	2.270	0.023*	2.002–27.334
HO	11.438	22.285	0.513	0.608	−32.239–55.115
R^2			0.712		
Adjusted R^2			0.682		
F			$F(18,176) = 24.204$, $p = 0.000$		
D-W value			2.018		

Implicit variable: RP; *$p < 0.05$ **$p < 0.01$.

3.6. Results of XGBoost regression

Table 10 demonstrates the setting of each model parameter, including the training ratio, and the rest of the indicators are the parameter values related to the model tuning parameters.

Table 10. Model parameter setting of XGBoost model.

Parameter name	Value
Data preprocessing	None
Training set ratio	0.8
Booster type	gbtree
Number of learners	100
Learning rate	0.1
Maximum tree depth	6
Sample Sampling Rate	1.0
Feature Sampling Rate	1.0
Smallest sub-node weight	1.0
Split gain threshold	0.0
L1 regularisation	0.0
L2 regularisation	1.0

Table 11 shows the importance of the contribution of each heading to the model

with a summed value of 1. PT weights 49.66%, the feature has the highest weight and plays a crucial role in model construction; SC weights 8.36%; ME weights 7.13%; TR weights 6.59%; the weight of the above four features together accounts for 71.74%; the remaining 14 headings GR, HO, DV, LI, PH, TB, PR, DE, NH, CP, CR, LA, BC, and BA were 3.55%, 3.42%, 3.20%, 3.16%, 2.30%, 2.08%, 2.07%, 1.73%, 1.43%, 1.35%, 1.25%, 1.12%, 0.99%, and 0.62%, respectively.

Table 11. Feature weight value of XGBoost regression (n = 195).

Items	Weights
LA	0.011
BA	0.006
LI	0.032
DV	0.032
DE	0.017
BC	0.010
PT	0.497
PH	0.023
GR	0.035
PR	0.021
CP	0.014
TB	0.021
CR	0.013
NH	0.014
ME	0.071
TR	0.066
SC	0.084
HO	0.034

Table 12. Fit results of XGBoost regression (n = 195).

Index	Clarification	Training set	Test set
R^2	The degree of fit indicator, the larger, the better between 0 and 1	1.000	0.541
Mean absolute error value (MAE)	L1 loss is the difference between the mean of the actual value and the fitted values' mean. The closer to 0, the better	22.111	1847.963
Mean Square Error (MSE)	L2 loss, the mean sum of squared errors. The closer to 0, the better	1073.128	5,743,765.687
Root Mean Square Error (RMSE)	MSE open root sign, average gap value	32.759	2396.615
Median absolute error (MAD)	The absolute value of the residuals of the predicted value from the median, independent of outliers, the smaller, the better	15.316	1287.547
Mean Absolute Percentage Error (MAPE)	Mean percentage error, independent of outliers, the smaller the better	0.003	0.091
Explainable Variance Score (EVS)	The measure of the strength of the model in explaining data fluctuations, between [0,1], the larger, the better	1.000	0.548
Root mean square logarithmic error (MSLE)	It penalizes underprediction more (less use) when RMSE is the same	0.000	0.037

The model evaluation results metrics in **Table 12** are used to evaluate the models' strengths and weaknesses and compare them; eight evaluation metrics are provided, including four metrics such as R^2, MAE, MSE, and RMSE, which are used more often. The fit of the metrics is significantly better in the training set than in the test set, implying an overfitting problem; the fit metrics are not abnormal (not within the standard range), and the R^2 values are all greater than zero [41].

3.7. Results of random forest regression

Table 13 shows the setting of each model parameter. The first parameter for data processing includes the training ratio, and the rest of the indicators are parameter values related to model tuning parameters.

Table 13. Parameter setting of the random forest model.

Parameter name	Value
Data preprocessing	None
Proportion of training set	0.8
Number of decision trees	100
Node splitting criterion	Squared_error
Minimum number of samples for node splitting	2
Minimum number of leaf node samples	1
Maximum tree depth	No limit
Maximum number of features	Auto
Whether to put back sampling	Yes
Whether or not out-of-bag data testing is performed	Yes
Data Preprocessing	None
Proportion of training set	0.8

Table 14 shows the significance of the contribution of each heading to the model with a summed value of 1. PT weights 61.43%, the feature has the highest weight and plays a crucial role in model construction; ME weights 4.79%; CR weights 3.99%; the combined weight of the above three features accounts for 70.22%; and the remaining 15 headings, TR, PH, SC, PR, HO, CP, BA, DV, LA, GR, NH, TB, LI, and DE, BC were 3.99%, 3.44%, 2.86%, 2.71%, 2.54%, 2.29%, 2.04%, 1.85%, 1.71%, 1.70%, 1.57%, 1.42%, 0.69%, 0.53%, and 0.45%, respectively.

Table 14. Feature weight value of random forest regression (n = 195).

Items	Weights
LA	0.017
BA	0.020
LI	0.007
DV	0.018
DE	0.005
BC	0.004

Table 14. (*Continued*).

Items	Weights
PT	0.614
PH	0.034
GR	0.017
PR	0.027
CP	0.023
TB	0.014
CR	0.040
NH	0.016
ME	0.048
TR	0.040
SC	0.029
HO	0.025

As illustrated in **Table 15**, The fact that the metrics fit significantly better in the training set of the random forest than in the test set means that there is also an overfitting problem if the fit metrics are not abnormal (not within the standard range), e.g., the R^2 value appears to be less than zero.

Table 15. Fit results of random forest regression (n = 195).

Index	Clarification	Training set	Test set
R^2	The degree of fit indicator, the larger, the better between 0 and 1	0.960	0.460
Mean absolute error value (MAE)	L1 loss is the difference between the mean of the actual value and the fitted values' mean. The closer to 0, the better	646.495	1611.784
Mean Square Error (MSE)	L2 loss, the mean sum of squared errors. The closer to 0, the better	736,769.896	4,332,247.126
Root Mean Square Error (RMSE)	MSE open root sign, average gap value	858.353	2081.405
Median absolute error (MAD)	The absolute value of the residuals of the predicted value from the median, independent of outliers, the smaller, the better	480.720	1776.960
Mean Absolute Percentage Error (MAPE)	Mean percentage error, independent of outliers, the smaller the better	0.085	0.060
Explainable Variance Score (EVS)	The measure of the strength of the model in explaining data fluctuations, between [0,1], the larger, the better	0.960	0.500
Root mean square logarithmic error (MSLE)	It penalizes underprediction more (less use) when RMSE is the same	0.005	0.035

3.8. Comparative analysis

Figure 5 shows the R^2 values of the regression algorithms applied to the house price prediction model.

OLS performs well in this case with the highest R^2 value. This indicates that the relationship between the predictor and target variables is reasonably linear and does not seriously violate the OLS assumptions (e.g., multicollinearity, homoskedasticity). Lasso regression (0.711) is ranked second. the R^2 value is very close to the OLS, suggesting that the model benefits from regularization and may be able to address

some of the multicollinearity issues without sacrificing too much predictive power. Robust regression (0.7) ranks third. The slightly lower R^2 suggests that there may be some outliers in the data that are not handled well by OLS but can be handled by robust regression, albeit with a slightly reduced fit for most of the data. Stepwise regression (0.687) ranks fourth. It has a marginally lower R^2 than the other linear models, which suggests that the stepwise method did not select the best set of predictor variables or that there was overfitting in the selection process. Ridge regression (0.65) came in fifth, with a lower R^2 value suggesting that while it can handle multicollinearity, it may over-penalize some of the coefficients, resulting in a slight reduction in fit. XGBoost regression (0.541) came in sixth. The lower R^2 indicates that XGBoost performs poorly as the linear model. This is mainly due to the smaller sample size (195), resulting in a more complex model that does not fit the training data well. The Random Forest regression (0.46) comes in last. the significantly lower R^2 suggests that the Random Forest may have overfitted the training data, especially with the smaller sample size. This indicates that the relationships in the data are not as complex as the RF model assumes.

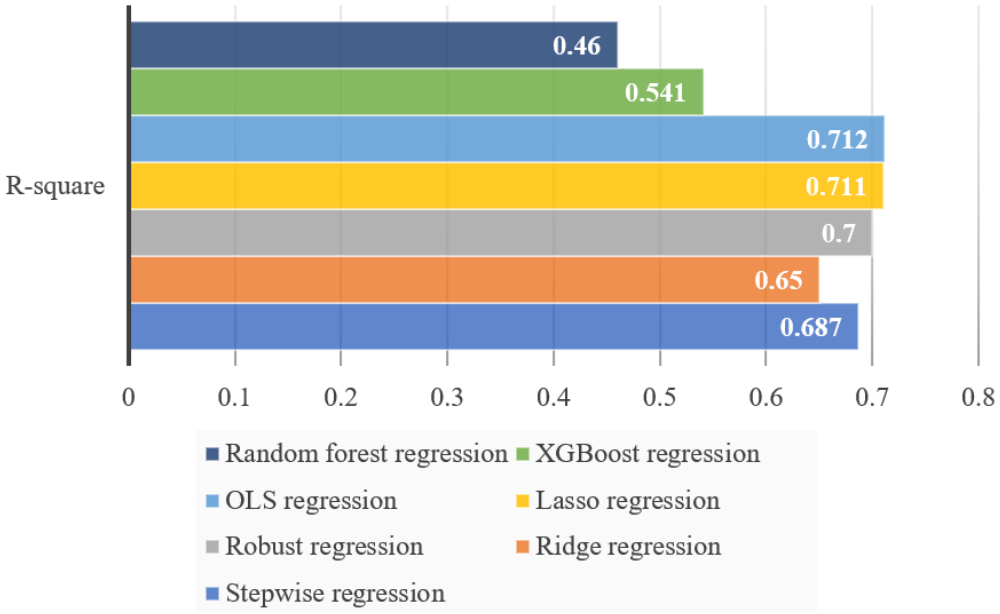

Figure 5. The heatmap of Pearson correlation analysis.

With only 195 samples, complex models such as XGBoost and Random Forests may perform poorly due to overfitting or insufficient data to capture complex patterns. Simple models such as OLS, Lasso, and Ridge are more straightforward and tend to perform better on smaller datasets. More complicated models, such as XGBoost and Random Forests, require a larger dataset to learn from the data effectively. Robust regression suggests that there are outliers that simple models may not handle well. Lower R^2 for stepwise regression may be due to poor variable selection, resulting in the model not capturing the complete variance explained by the predictors.

Future research suggests increasing the dataset size to improve the performance of more complex models. Secondly, it should be ensured that the features used in the model are appropriately scaled and transformed when needed. Third, perform more sophisticated hyperparameter tuning, especially for complex models like XGBoost

and Random Forest, to find the optimal settings. Finally, it is also recommended that similar studies use cross-validation to assess model performance more robustly, which helps to understand the true predictive power of the model.

4. Discussion

4.1. Promoting the renewal of residential district design

Future residential communities will be intelligent, ecological, and community-based complexes [42]. They will incorporate advanced technologies such as smart home systems, renewable energy, and intelligent transport [43]. By sensing and comprehending the living habits of the occupants, they will provide a quality living experience. Equipment such as smart parking, security, remote control of water, electricity, and coal, and intelligent regulation of temperature and humidity will become standard [44,45]. We also suggest paying more attention to the ecological environment, including green coverage, rainwater collection and utilization, and waste separation and disposal, to create a livable eco-community. Incorporating environmental and gardening elements will help satisfy the desire of residents to be close to nature. For example, the "White Tree" project in Montpellier, France, has sky gardens and private balconies for each household, blending in with the surrounding environment. Current construction technology allows humans to live anywhere on the planet, regardless of the terrain. An example is the "Cloud Corridor" project in Los Angeles, USA, a steel building constructed in a forest. In addition, with reasonable property fees, community management should emphasize shared facilities and social spaces, encouraging residents to help each other and participate in community activities, thus forming a more harmonious and interactive community relationship. Such activities that enhance the community's sense of well-being, access, and honor will also promote tremendous enthusiasm among homebuyers [46].

4.2. Encouraging stimulus intervention by the government

The government can support the development of the property market in several ways. The first is to formulate and adjust relevant policies, including real estate taxation policies, land supply policies, and health and education policies, to promote the smooth development of the market and a balance between supply and demand [47]. For example, the reserve requirement ratio for deposits influences the market's capital flow and stimulates the demand for home purchases. Some home purchase subsidies and preferential policies, such as tax concessions for first-time home buyers and preferential interest rates for provident fund loans, can also greatly encourage residents to purchase homes. Many cities have launched preferential policies to support the purchase of homes by families with many children, adding 200,000 to 300,000 yuan to the original actual loan amount; the amount can be further relaxed for high-level talents, such as doctors, and appropriate amounts of talent subsidies are given. Such a diversified housing policy is also conducive to promoting urban home ownership among the rural population, especially migrant workers who are contracted to cultivate farmland but do not have a home base in their hometowns [48,49]. Second, based on the study's findings, we believe that the reform of property charges is likewise a

complex and critical topic involving the rights and interests of property owners, service quality, and transparency of charges. Some regions in China have started implementing government-guided pricing to ensure that property fees are not too high. For example, Chongqing Municipality has stipulated that all residential pre-property service charges are to be managed at government-guided prices. Secondly, charges for services exceeding the maximum grade should be strictly identified as over-graded service programs, and the relevant authorities should approve charges. Local governments must formulate a policy on property service charges for undecorated and unused residential properties to ensure fairness and reasonableness. Owners' committees should work with property service enterprises to set reasonable prices and decide on property fee adjustment programs through voting at owners' meetings.

4.3. Accelerating the process of retrofitting old buildings

The number of old residential districts in China is currently considerable. About 160,000 old residential districts were built before 2000, involving more than 42 million households and a total floor area of about 4 billion square meters [50]. Although, in 2023, the country will start renovating 53,700 old urban neighborhoods, benefiting 8.97 million households, the aging of buildings continues, and there is still a massive gap in the market [51]. Since China's real estate reform in 1998, the neighborhoods previously constructed through the welfare housing system have long been out of step with the times regarding the environment, building facades, pipelines and ducts, security measures, and neighborhood amenities. Many old buildings are facing aging functions, waste of resources, and environmental problems, and by renovating these buildings, urban space can be effectively utilized to enhance the efficiency and value of building use. It is worth worrying that if we do not accelerate the pace of renovation of old buildings, the problems faced in the future will be even more severe. This is because the maintenance costs of high-rise buildings after 2000 are greater than those of low-rise residential buildings before 2000. Although China has established an overhaul fund, it can be utilized after the housing warranty period. With inflation and other factors, it will inevitably be significantly reduced twenty years later. In the United States, for example, where high-rise buildings were developed earlier, many middle-class people have moved out of high-rise housing. In addition, the renovated space's openness, rhythm, and richness of colors will affect people's moods, and the renovation of more traditional ethnic-style buildings will also help improve the city's aesthetic fatigue [52].

5. Conclusion

The decision to purchase residential property in China is critical in the lives of most adult citizens. Therefore, real estate appraisals and forecasts can provide helpful information to help sellers make effective decisions and facilitate fair real estate transactions. Real estate prices in the same city vary depending on the basic parameters of various properties. This study uses machine learning models, including stepwise regression, OLS regression, Lasso regression, random forest, and XGBoost regression, to predict real estate transaction prices from actual transaction data in Changsha. The empirical results show that the machine learning models in this study are all suitable

for predicting real estate prices, but there are differences in their effectiveness. Among them, OLS and Lasso regression outperform other prediction models and obtain more accurate results in terms of R^2 than in some previous studies.

It is worth noting that this study also has limitations. Firstly, the study only uses data from February 2024 in the Changsha area, which provides consistent evidence of time and hinders the extension of the predictive effect of time. Second, this study focuses only on the residential market, so the results do not apply to other property types, such as shops. This study suggests that future research could expand data types, such as developer regulatory account balances, social media promotional efforts, google map satellite imagery, and regional purchasing power metrics, as sources of input to the ML model to improve predictive accuracy. Also, more deep learning neural networks can be considered for more attempts in the future.

Author contributions: Conceptualization, YJ and AHA; methodology, YJ; software, YJ; validation, AHA, NAH and NAB; formal analysis, YJ; investigation, YJ; resources, YJ; data curation, YJ; writing—original draft preparation, YJ; writing—review and editing, YJ; visualization, YJ; supervision, AHA, NAH and NAB; project administration, YJ. All authors have read and agreed to the published version of the manuscript.

Acknowledgments: The authors thank the Hong Kong Shue Yan University academics for guiding this study.

Conflict of interest: The authors declare no conflict of interest.

References

1. Feng Y, Wahab MA, Azmi NAB, et al. Chinese Residents' Willingness to Buy Housing: An Evaluation in Nanyang City, Henan Province, China Based on the Extension Cloud Model. Buildings. 2022; 12(10): 1695. doi: 10.3390/buildings12101695
2. National Bureau of Statistics. Statistical Bulletin of the People's Republic of China on National Economic and Social Development, 2023. Available online: https://www.stats.gov.cn/sj/zxfb/202402/t20240228_1947915.html (accessed on 22 June 2024).
3. Li B, Li RYM, Wareewanich T. Factors Influencing Large Real Estate Companies' Competitiveness: A Sustainable Development Perspective. Land. 2021; 10(11): 1239. doi: 10.3390/land10111239
4. Li Y, Xiang Z, Xiong T. The Behavioral Mechanism and Forecasting of Beijing Housing Prices from a Multiscale Perspective. Discrete Dynamics in Nature and Society. 2020; 2020: 1-13. doi: 10.1155/2020/5375206
5. Rico-Juan JR, Taltavull de La Paz P. Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain. Expert Systems with Applications. 2021; 171: 114590. doi: 10.1016/j.eswa.2021.114590
6. Xu X, Zhang Y. House price forecasting with neural networks. Intelligent Systems with Applications. 2021; 12: 200052. doi: 10.1016/j.iswa.2021.200052
7. Li R, Li H. Have Housing Prices Gone with the Smelly Wind? Big Data Analysis on Landfill in Hong Kong. Sustainability. 2018; 10(2): 341. doi: 10.3390/su10020341
8. Miles D, Monro V. UK house prices and three decades of decline in the risk-free real interest rate. Economic Policy. 2021; 36(108): 627-684. doi: 10.1093/epolic/eiab006
9. Duca JV, Muellbauer J, Murphy A. What Drives House Price Cycles? International Experience and Policy Issues. Journal of Economic Literature. 2021; 59(3): 773-864. doi: 10.1257/jel.20201325
10. Barron K, Kung E, Proserpio D. The Effect of Home-Sharing on House Prices and Rents: Evidence from Airbnb. Marketing Science. 2021; 40(1): 23-47. doi: 10.1287/mksc.2020.1227

11. Bangura M, Lee CL. House price diffusion of housing submarkets in Greater Sydney. Housing Studies. 2019; 35(6): 1110-1141. doi: 10.1080/02673037.2019.1648772
12. Liu G. Research on Prediction and Analysis of Real Estate Market Based on the Multiple Linear Regression Model. Scientific Programming. 2022; 2022: 1-8. doi: 10.1155/2022/5750354
13. Madhuri CHR, Anuradha G, Pujitha MV. House Price Prediction Using Regression Techniques: A Comparative Study. In: Proceedings of the 2019 International Conference on Smart Structures and Systems (ICSSS). doi: 10.1109/icsss.2019.8882834
14. Kim J, Lee Y, Lee MH, et al. A Comparative Study of Machine Learning and Spatial Interpolation Methods for Predicting House Prices. Sustainability. 2022; 14(15): 9056. doi: 10.3390/su14159056
15. Thamarai M, Malarvizhi SP. House Price Prediction Modeling Using Machine Learning. International Journal of Information Engineering and Electronic Business. 2020; 12(2): 15-20. doi: 10.5815/ijieeb.2020.02.03
16. Qin L, Zong W, Peng K, et al. Assessing Spatial Heterogeneity in Urban Park Vitality for a Sustainable Built Environment: A Case Study of Changsha. Land. 2024; 13(4): 480. doi: 10.3390/land13040480
17. Zhou Z, Yang F, Li J, et al. Identification of Critical Areas of Openness-Vitality Intensity Imbalance in Waterfront Spaces and Prioritization of Interventions: A Case Study of Xiangjiang River in Changsha, China. Land. 2024; 13(5): 686. doi: 10.3390/land13050686
18. Anjuke. Changsha New Homes Information. Available online: https://m.anjuke.com/cs/ (accessed on 13 March 2024).
19. Li N, Li RYM, Nuttapong J. Factors affect the housing prices in China: a systematic review of papers indexed in Chinese Science Citation Database. Property Management. 2022; 40(5): 780-796. doi: 10.1108/pm-11-2020-0078
20. Liu M, Ma QP. Determinants of house prices in China: a panel-corrected regression approach. The Annals of Regional Science. 2021; 67(1): 47-72. doi: 10.1007/s00168-020-01040-z
21. Wang Z, Feng Y, Li Y, et al. Inheritance dynamics and housing price fluctuations: Evidence from the China household finance survey. Finance Research Letters. 2024; 67: 105743. doi: 10.1016/j.frl.2024.105743
22. Sun Q, Javeed SA, Tang Y, et al. The impact of housing prices and land financing on economic growth: Evidence from Chinese 277 cities at the prefecture level and above. PLOS ONE. 2024; 19(4): e0302631. doi: 10.1371/journal.pone.0302631
23. Papazafeiropoulos G. Stepwise Regression for Increasing the Predictive Accuracy of Artificial Neural Networks: Applications in Benchmark and Advanced Problems. Modelling. 2024; 5(1): 153-179. doi: 10.3390/modelling5010009
24. Ma L, Yang H, Yang J. A Multimodal Teaching Quality Evaluation for Hybrid Education Based on Stepwise Regression Analysis. Journal on special topics in mobile networks and applications/Mobile networks and applications. 2023; 1-11.
25. Arashi M, Roozbeh M, Hamzah NA, et al. Ridge regression and its applications in genetic studies. PLOS ONE. 2021; 16(4): e0245376. doi: 10.1371/journal.pone.0245376
26. Hoerl RW. Ridge Regression: A Historical Context. Technometrics. 2020; 62(4): 420-425. doi: 10.1080/00401706.2020.1742207
27. Samaniego A. CAPM-alpha estimation with robust regression vs. linear regression. Análisis Económico. 2023; 38(97): 27-37. doi: 10.24275/uam/azc/dcsh/ae/2022v38n97/samaniego
28. Gao C. Robust regression via mutivariate regression depth. Bernoulli. 2020; 26(2). doi: 10.3150/19-bej1144
29. Verardi V, Croux C. Robust Regression in Stata. SSRN Electronic Journal. 2008. doi: 10.2139/ssrn.1369144
30. Xin SJ, Khalid K. Modelling House Price Using Ridge Regression and Lasso Regression. International Journal of Engineering & Technology. 2018; 7(4.30): 498. doi: 10.14419/ijet.v7i4.30.22378
31. Roth V. The Generalized LASSO. IEEE Transactions on Neural Networks. 2004; 15(1): 16-28. doi: 10.1109/tnn.2003.809398
32. Sanchez JM. Estimating Detection Limits in Chromatography from Calibration Data: Ordinary Least Squares Regression vs. Weighted Least Squares. Separations. 2018; 5(4): 49. doi: 10.3390/separations5040049
33. Nascimento RS, Froes RES, e Silva NOC, et al. Comparison between ordinary least squares regression and weighted least squares regression in the calibration of metals present in human milk determined by ICP-OES. Talanta. 2010; 80(3): 1102-1109. doi: 10.1016/j.talanta.2009.08.043
34. Zhang X, Yan C, Gao C, et al. Predicting Missing Values in Medical Data Via XGBoost Regression. Journal of Healthcare Informatics Research. 2020; 4(4): 383-394. doi: 10.1007/s41666-020-00077-1
35. Shehadeh A, Alshboul O, Al Mamlook RE, et al. Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression. Automation in

Construction. 2021; 129: 103827. doi: 10.1016/j.autcon.2021.103827
36. Iannace G, Ciaburro G, Trematerra A. Wind Turbine Noise Prediction Using Random Forest Regression. Machines. 2019; 7(4): 69. doi: 10.3390/machines7040069
37. Mendez G, Lohr S. Estimating residual variance in random forest regression. Computational Statistics & Data Analysis. 2011; 55(11): 2937-2950. doi: 10.1016/j.csda.2011.04.022
38. Yao Q, Li RYM, Song L, et al. Construction safety knowledge sharing on Twitter: A social network analysis. Safety Science. 2021; 143: 105411. doi: 10.1016/j.ssci.2021.105411
39. Daoud JI. Multicollinearity and Regression Analysis. Journal of Physics: Conference Series. 2017; 949: 012009. doi: 10.1088/1742-6596/949/1/012009
40. Tiku ML. Tables of the Power of the F-Test. Journal of the American Statistical Association. 1967; 62(318): 525. doi: 10.2307/2283980
41. Colin Cameron A, Windmeijer FAG. An R-squared measure of goodness of fit for some common nonlinear regression models. Journal of Econometrics. 1997; 77(2): 329-342. doi: 10.1016/S0304-4076(96)01818-0
42. Mao Q, Wang L, Guo Q, et al. Evaluating Cultural Ecosystem Services of Urban Residential Green Spaces from the Perspective of Residents' Satisfaction with Green Space. Frontiers in Public Health. 2020; 8. doi: 10.3389/fpubh.2020.00226
43. Feng Q, Wang Y, Chen C, et al. Effect of Homebuyer Comment on Green Housing Purchase Intention—Mediation Role of Psychological Distance. Frontiers in Psychology. 2021; 12. doi: 10.3389/fpsyg.2021.568451
44. Guo M, Xiao S. An empirical analysis of the factors driving customers' purchase intention of green smart home products. Frontiers in Psychology. 2023; 14. doi: 10.3389/fpsyg.2023.1272889
45. Bai S, Li F, Xie W. Green but Unpopular? Analysis on Purchase Intention of Heat Pump Water Heaters in China. Energies. 2022; 15(7): 2464. doi: 10.3390/en15072464
46. Zhao S, Chen L. Exploring Residents' Purchase Intention of Green Housings in China: An Extended Perspective of Perceived Value. International Journal of Environmental Research and Public Health. 2021; 18(8): 4074. doi: 10.3390/ijerph18084074
47. Ma D, Lv B, Li X, et al. Heterogeneous Impacts of Policy Sentiment with Different Themes on Real Estate Market: Evidence from China. Sustainability. 2023; 15(2): 1690. doi: 10.3390/su15021690
48. Song Y, Zhang C. City size and housing purchase intention: Evidence from rural-urban migrants in China. Urban Studies. 2019; 57(9): 1866-1886. doi: 10.1177/0042098019856822
49. Zou J, Chen J, Chen Y. Hometown landholdings and rural migrants' integration intention: The case of urban China. Land Use Policy. 2022; 121: 106307. doi: 10.1016/j.landusepol.2022.106307
50. Xiaolan Z. 160,000 old neighborhoods look forward to a 'new look'. People's Daily Online. 2019. Available online: https://house.people.com.cn/n1/2019/0726/c164220-31257403.html (accessed on 23 June 2024).
51. Urban Construction Division (UCD). Nationwide, 53,700 new urban old districts to be renovated by 2023. Available online: https://www.mohurd.gov.cn/xinwen/gzdt/202402/20240201_776526.html (accessed on 23 June 2024).
52. Zeng L, Li RYM, Li R. Chromaticity Analysis on Ethnic Minority Color Landscape Culture in Tibetan Area: A Semantic Differential Approach. Applied Sciences. 2024; 14(11): 4672. doi: 10.3390/app14114672