Urban Housing Price Prediction and Service Level Evaluation Based on Machine Learning

Jiaye Chen [1, #], Jiarui Huang [2, #], Huawei Huang [3], Ali Zulfiqar [4, *]

[1] Business School, University of Southampton, Southampton, United Kingdom

[2] School of mathematics and statistics, Nanjing University of Science and Technology, Nanjing, China

[3] School of Social Science, The University of Manchester, Manchester, United Kingdom

[4] Department of Mechanical Engineering, The University of Hong Kong, Hong Kong China

* Corresponding Author Email: zulfiali@hku.hk

[#]These authors contributed equally

Abstract. With the acceleration of urbanization, the sustainable development of cities and the improvement of residents' quality of life have become major challenges facing the world. Taking Wenzhou City and Hohhot City of China as examples, this paper uses machine learning method to build urban housing price prediction and service level evaluation models. First, the stochastic forest regression and ridge regression models are used to predict the housing price, and the forecasting effect of the two models is compared. The results show that the stochastic forest regression model has better performance in capturing nonlinear relationship and forecasting accuracy. Secondly, a multi-index comprehensive evaluation model is constructed to evaluate the urban service level, and the factors affecting the service level are determined by weighted summation method. The results show that optimizing the greening rate and property management fees can effectively enhance the service level and improve the quality of life of residents. The research results of this paper provide a scientific basis for urban planning and decision-making, and provide a reference for the sustainable development of other cities. Future studies can further explore other factors that affect housing prices and service levels, and incorporate more advanced machine learning models to improve the predictive accuracy and practicality of the models.

Keywords: Random Forest Regression (RF), Ridge Regression (LR), service level assessment, multi-indicator comprehensive evaluation, urban sustainable development.

1. Introduction

Sustainable urban development and the improvement of residents' quality of life are major challenges facing the world. Urban housing price forecast and service level assessment are important bases for urban planning and decision-making [1]. In recent years, machine learning technology has made remarkable achievements in various fields, and its application in urban housing price prediction and service level assessment has gradually attracted attention.

At present, many scholars have used machine learning methods to forecast urban housing prices, and achieved good results [2,3]. In terms of service level evaluation, some scholars have also used machine learning methods to build evaluation models [4,5].

Taking Wenzhou City and Hohhot city as examples, this paper uses random forest regression and ridge regression models to predict housing prices, and constructs a multi-index comprehensive evaluation model to evaluate service level, aiming to provide scientific basis for urban planning and decision-making, and provide reference for sustainable development of other cities.

2. Materials and methods

2.1. Data

This study collected the data of housing price, green rate, plot ratio, property management fee and so on in Wenzhou and Hohhot through field investigation.

In the data preprocessing stage, the missing value is processed first, and the integrity of the data is ensured by means of filling or deleting the mean value. The data are standardized to make the features of different magnitudes comparable. For statistical analysis, descriptive statistics are calculated for each property characteristic to understand its basic distribution characteristics.

2.2. Method introduction

2.2.1 Random forest regression model introduction

(1) Principle

Random forest regression is an ensemble learning algorithm based on decision tree algorithm. The core idea is to combine multiple decision trees to form a "forest" and make final predictions through the collective decisions of these decision trees [6]. Specifically, the building process consists of the following key steps:

1) Random selection of samples: Random Sampling with replacement (called Bootstrap Sampling) is performed from the original data set to create multiple different training subsets [7]. This means that some samples may appear in multiple subsets, while others may not be selected, and about a third of the samples may not appear in any one subset, and these unselected samples can be used for subsequent model evaluation, called Out-Of-Bag data (OOB).

2) Random selection of features: When constructing each decision tree, for each node split, a part of the features is randomly selected from all the features, and then the best split feature is selected from these randomly selected features. This randomness helps to make each tree unique, reducing the variance of the model.

3) Decision tree construction and integration: Using selected samples and features, the decision tree is grown according to the construction rules of the decision tree (such as using indicators such as Gini index or information gain). Finally, for regression problems, the prediction result of the random forest is the average of the prediction results of all decision trees [8].

(2) Advantages

1) Ability to process high-dimensional data: Random Forest can well process data with a large number of features, and its performance will not be significantly reduced due to the increase of data dimensions, because in the feature selection stage, some features are randomly selected for splitting, which avoids dimension disaster to a certain extent.

2) Handling of nonlinear relationships: Unlike linear regression and other models, random forest can capture nonlinear relationships in data well, because the decision tree itself can model complex nonlinear data, and the integration of multiple decision trees further enhances the learning ability of nonlinear patterns.

3) Strong robustness: It is relatively insensitive to noise and outliers in the data, because the influence of outliers is dispersed across different decision trees when building decision trees, and random sampling also reduces the impact of individual outliers on the overall model.

4) Evaluable feature importance: An estimate of feature importance can be given by calculating the contribution of each feature to model performance when building the decision tree, which is very helpful for feature selection and understanding data. For example, you can use the feature_importances_ attribute to see the feature importance.

2.2.2 Ridge regression model introduction

Ridge regression model is an improvement of ordinary linear regression. Its basic principle is to add L2 regularization term to the objective function of traditional least square method. The objective function is

$$J(\beta) = ||y - X\beta||^2 + \alpha||\beta||^2 \quad (1)$$

where y represents the dependent variable. X represents the independent variable matrix, β is the regression coefficient vector, α is the regularization parameter [9].

When minimizing this objective function, ridge regression not only seeks to minimize the error between the predicted value and the actual value like ordinary linear regression, but also restricts the size of the regression coefficient β by the term $\alpha||\beta||^2$. This constraint brings advantages in dealing with multicollinearity problems: when there is multicollinearity between the independent variables, that is, there is a high linear correlation between them, the least square estimation of ordinary linear regression will cause the variance of the regression coefficient to increase, and the estimation will become unstable. However, due to the addition of regularization term in ridge regression, the regression coefficient shrinks to zero. Even if there is a strong correlation between independent variables, some coefficients will not be too large or too small, thus reducing the excessive dependence of the model on collinearity features and improving the stability of the model. At the same time, by adjusting the value of α, it can balance the bias and variance, sacrifice the bias to a certain extent, but significantly reduce the variance, so as to improve the prediction performance of the model in the presence of multicollinearity, make the model more robust and more suitable for data sets with multicollinearity [10].

3. Results

3.1. Construction and analysis of urban housing price prediction model based on random forest and ridge regression

3.1.1 Establishment of Random Forest Regression (RF) and Ridge Regression (LR) Models

We selected two different models to predict property prices: the Random Forest Regression (RF) model and the Ridge Regression (LR) model. The basic structure of each model is described below.

The Random Forest Regression model is a regression method based on ensemble learning. It obtains the final prediction by averaging the results of multiple decision trees. The predictive formula is as follows:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t(X) \quad (2)$$

where T represents the number of decision trees, ft (X)denotes the t tree prediction. Random Forest improves prediction accuracy by combining the votes and averages of multiple trees, making it well-suited for capturing complex non-linear relationships between features.

Ridge Regression, on the other hand, is a linear regression method with L2 regularization, which effectively mitigates overfitting caused by multicollinearity among features.

The objective function for Ridge Regression is:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \alpha \sum_{j=1}^{p} \beta_j^2 \quad (3)$$

Where α is the regularization coefficient, which controls the weight of the regularization term, and β represents the model coefficients. By adding the L2 penalty term, Ridge Regression helps reduce model complexity and improve generalization ability.

Since the features in the model have different units and magnitudes, we standardize the features to eliminate these effects:

$$X_{scaled} = \frac{X-\mu}{\sigma} \quad (4)$$

Where μ is the mean of the feature, σis the standard deviation of the feature.

Assuming that future features (such as greenery rate, floor area ratio, and property management fees) change according to a specific annual growth rate, we predict the trend of these features over

the next 10 years. If the feature X future has an annual growth rate of r,the value of the feature in the t year can be expressed as:

$$X_{future} = X_{current} \times (1 + r)^t \quad (5)$$

3.1.2 Model Solution

Based on the model design outlined above, the process of data loading, feature engineering, model training, and prediction is implemented using Python code.

(1) Data Loading and Preprocessing: The property data for City 1 and City 2 is loaded and cleaned. Features such as greenery rate, floor area ratio, and property management fees are extracted. Missing values are handled, and standardization is performed.

(2) Model Training: The data is divided into training and testing sets. Both the Random Forest Regression model and the Ridge Regression model are trained using the respective datasets.

(3) Model Evaluation: The performance of the models is evaluated using the Mean Absolute Error (MAE), with the following formula:

$$\text{MAE} \cdot = \cdot \frac{1}{n}\sum_{i=1}^{n} \cdot |y_{1\cdots} - \cdot \hat{y}_i| \quad (6)$$

(4) Future Property Price Prediction: Based on the simulated results of future features, the future property price trends for the next 10 years are predicted.

3.1.3 Consequence of Model Prediction

In Figure 1, the MAE of the random forest model is lower than that of the ridge regression model for both Wenzhou and Hohhot, indicating that the random forest model performs slightly better in predicting in these two cities. This may be because the random forest model can better capture the nonlinear relationships in housing price data, whereas the ridge regression model is primarily suitable for datasets with more obvious linear relationships.

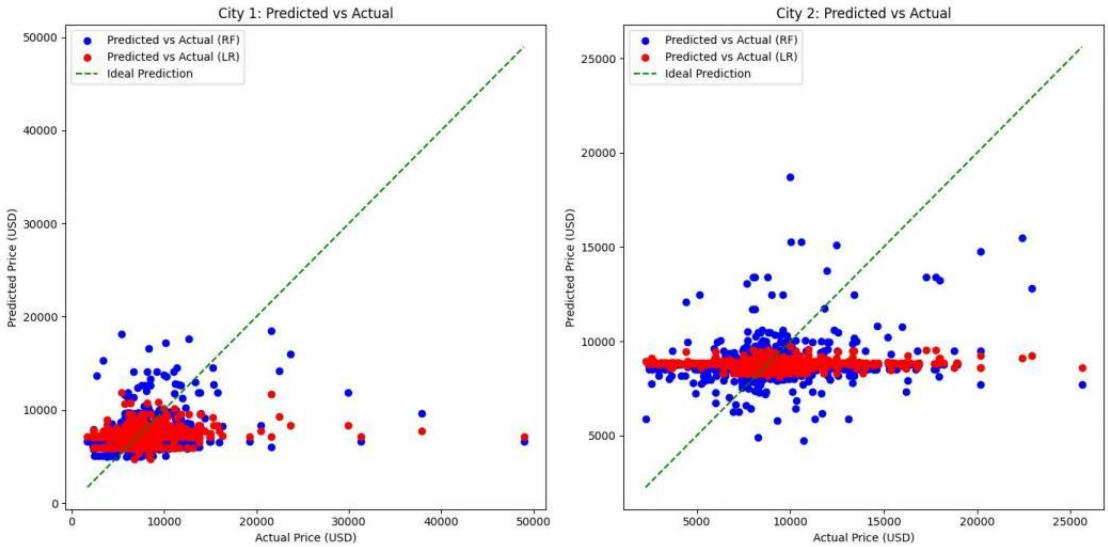

Figure 1. The predicted and true value of housing prices in the two cities

Figure 1 illustrates the comparison between predicted and actual property prices in Wenzhou and Hohhot. Blue points represent the predictions from the Random Forest model, red points represent those from the Ridge Regression model, and the green dashed line represents the ideal prediction line (i.e., y = x). From the comparison, the following conclusions can be drawn.

Most prediction points are concentrated in the lower price range (below 20,000 USD) and are relatively densely distributed, indicating that the models exhibit smaller errors when predicting mid-to-low-priced properties.

When property prices exceed 20,000 USD, the deviation between predicted and actual values becomes noticeably larger. This is particularly evident for some high-priced properties, where the prediction errors are more significant.

Reasons may include:

(1) Market Dynamics: Low-priced properties are more stable, while high- priced properties experience more volatility due to economic and policy changes.

(2) Feature Impact: Location, size, and fees predict low-priced proper- ties well, but high-priced properties often have unique, harder-to-capture features.

(3) School District Influence: Proximity to high-quality schools affects prices significantly, making predictions complex due to fluctuating demand.

(4) Property Types: High-priced properties, like luxury homes, have unique factors (e.g., architectural style, amenities) that standard models may not fully capture.

3.2. Urban service level evaluation model based on multi-index comprehensive evaluation

3.2.1 Development of the Service Level Model

(1) The data cleaning part mainly includes the following steps:

Handling Missing Values:Removing or imputing missing values in property prices and other key features to ensure data completeness.

Processing Range Data:For features such as green coverage rate, floor area ratio, and property management fees that might include ranges, the median of the range is used as a representative value. Specifically, a function is defined to parse these range values into their averages:

$$\text{Range} \cdot \text{Mean} \cdot = \cdot \frac{\text{Start Value} + \text{EndValue}}{2} \tag{7}$$

(2) In the feature selection, we selected the following variables to construct the feature matrix X:

Greening Rate: Represents environmental quality.

Floor area ratio: Reflects building density.

Property management fee: Represents the cost of daily maintenance.

(3) Statistical Description: For each city, we calculated descriptive statistics for each indicator, such as mean, median, and standard deviation.

(4) Comparative Analysis:

Used boxplots and histograms to observe the distribution of each indicator.

Compared the mean values of GR, FAR, and PMF to identify the strengths and weaknesses of each city.

3.2.2 Problem Solving

(1) By calculating the descriptive statistics (e.g., mean, standard deviation, and quartiles) for service-level indicators in each city, a foundation for subsequent analysis is established. For example, the formula for descriptive statistics of a specific indicator is as follows:

$$\text{Mean} \cdot = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{8}$$

where xi represents the indicator value for each sample,n represents the total number of samples.

(2)Data Distribution Visualization:

This study presents the distribution of service-level indicators for Wenzhou and Hohhot through histograms and kernel density estimation (KDE) plots. These visualizations intuitively illustrate the central tendency and distribution patterns of the indicators. The purpose of these visualizations is to examine whether the data distribution exhibits skewness, kurtosis, or other notable characteristics, such as significant differences in living costs between the two cities.

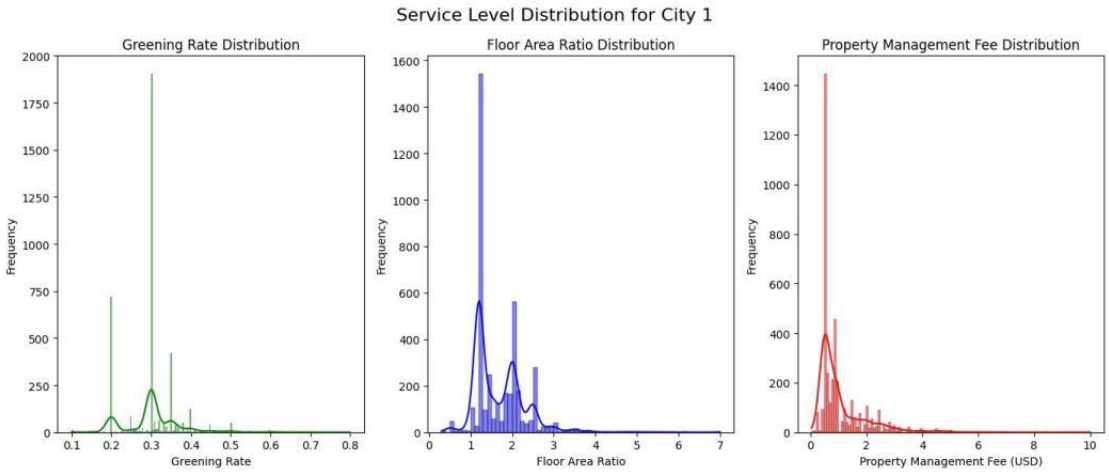

Figure 2. Service level of Wenzhou

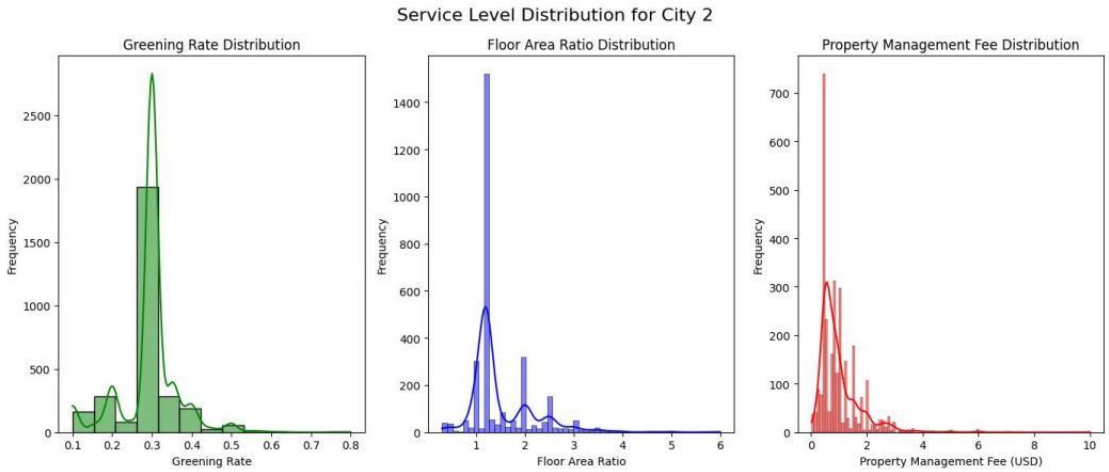

Figure 3. Service level of Hohhot

Summary Statistics for City 1:

	Greening rate	Floor area ratio	Property management fee (/m²/month USD)
count	4016.000000	4015.000000	3983.000000
mean	0.302040	1.672538	1.068411
std	0.070276	0.610209	0.902018
min	0.100000	0.300000	0.040000
25%	0.300000	1.200000	0.500000
50%	0.300000	1.500000	0.720000
75%	0.330000	2.000000	1.200000
max	0.800000	7.000000	10.000000

Summary Statistics for City 2:

	Greening rate	Floor area ratio	Property management fee (/m²/month USD)
count	3027.000000	3030.000000	2974.000000
mean	0.295389	1.503495	0.990551
std	0.073149	0.679704	0.748002
min	0.100000	0.300000	0.020000
25%	0.300000	1.200000	0.500000
50%	0.300000	1.200000	0.800000
75%	0.300000	1.990000	1.200000
max	0.800000	6.000000	10.000000

Figure 4. Comparison of service level in two cities

Through a visual analysis of the distribution of service level indicators (Figure 4)— green space ratio, floor area ratio, and property management fee—in Wenzhou and Hohhot, the following conclusions were drawn:

1)From Figure 2 and Figure 3,The green space ratio in Wenzhou and Hohhot is concentrated around 0.3.

The average green space ratio in Wenzhou is 0.302, while in Hohhot it is 0.295, with a small difference.

The distribution of green space ratios in both cities is relatively concentrated, with a small standard deviation, indicating that the greening levels in most areas tend to be consistent.

2)In terms of floor area ratio, the average in Wenzhou is 1.67, higher than Hohhot's 1.50.

The floor area ratio distribution in Wenzhou is relatively concentrated, with a small standard deviation, indicating a denser development layout.

The floor area ratio distribution in Hohhot is slightly more dispersed, suggesting the presence of more low floor area ratio areas, which represent more spacious building layouts.

3)In terms of property management fees, the average fee in Wenzhou is 1.068 USD/month/square meter, while in Hohhot it is 0.991 USD/month/square meter. The maximum property management fee for high-end properties in both cities is 10 USD, indicating similarity between the two cities in terms of high-end property management fees.

Living costs in Wenzhou are relatively high.

Hohhot has lower management fees for low-tier properties, indicating a lower overall cost of living.

(3)Evaluating the Advantages and Disadvantages of Each City Based on Mean Comparison of Descriptive Statistics:

Wenzhou reached higher greening rate and floor area ratio may suggest better environmental quality and denser urban development.

Hohhot's lower property management fees might imply a more affordable cost of living.

4. Conclusions

Through in-depth analysis of urban housing price prediction and service level assessment, this study successfully applied random forest regression and ridge regression models to improve the accuracy of housing price prediction, especially in the prediction of low and medium price real estate. The construction of multi-criteria service level model also provides the direction for the improvement of urban service quality. Future research can further expand the influencing factors, such as considering the impact of geographical location and surrounding supporting facilities on housing prices, and improving the service level model by increasing indicators such as the convenience of public transportation and educational resources, so as to provide more accurate and comprehensive support for urban planning and management and promote sustainable urban development.

References

[1] Anders H, Ida S, Einar D S, et al. Locally interpretable tree boosting: An application to house price prediction[J]. Decision Support Systems,2024,178114106-.

[2] Fatemeh M, Vedat T, Basri H B, et al.Multiedge Graph Convolutional Network for House Price Prediction[J].Journal of Construction Engineering and Management,2023,149(11).

[3] Hanxiang Z, Yansong L, Paula B. Describe the house and I will tell you the price: House price prediction with textual description data[J]. Natural Language Engineering,2023,30(4):661-695.

[4] Yansong L, Paula B, Hanxiang Z. Imbalanced Multimodal Attention-Based System for Multiclass House Price Prediction[J]. Mathematics,2022,11(1):113-113.

[5] Geerts M, Broucke V S, Weerdt D J. Graph neural networks for house price prediction: do or don't?[J]. International Journal of Data Science and Analytics,2024, (prepublish):1-31.

[6] Chen S, Huang D, Yu S, et al.Developing a rapid COD detection method based on the fusion strategy of multi-depth hyperspectral data[J].Biochemical Engineering Journal,2025,215109630-109630.

[7] Du J, Jacinthe A P, Song K, et al.Maize crop residue cover mapping using Sentinel-2 MSI data and random forest algorithms[J].International Soil and Water Conservation Research,2025,13(1):189-202.

[8] Khosravi V, Gholizadeh A, Kodešová R, et al.Visible, near-infrared, and shortwave-infrared spectra as an input variable for digital mapping of soil organic carbon[J].International Soil and Water Conservation Research,2025,13(1):203-214.

[9] Ulrych U, Vasiljević N. Global currency hedging with ambiguity[J]. Journal of Banking and Finance,2025,172107366-107366.

[10] Hadil H, Khaled E, ErnestJohn I, et al.Comparison of Machine-Learning Algorithms for Estimating Cost of Conventional and Accelerated Bridge Construction Methods during Early Design Phase[J].Journal of Construction Engineering and Management,2025,151(3):