Prediction and Analysis of Rental Price using Random Forest Machine Learning Technique

Take Shanghai and Wuhan for example

Chuang Hu[1*, †], Rui Huang[2a†], Haijian Li[3b†]

[1]School of Transportation and Logistics Engineering, Wuhan University of Technology Wuhan, Hubei, 430070, China
[2]School of economics, Hunan Agricultural University Changsha, Hunan, 410125, China
[3]School of Statistics and Mathematics, Zhongnan University of Economics and Law Wuhan, Hubei,430073, China
[*]311923@whut.edu.cn
[a]1378711499@qq.com
[b]1106146687@qq.com
[†]These authors contributed equally.

Abstract

With the rapid development of China's real estate market, the real estate industry has become a significant part of Chinese national economy. However, the high housing prices in the first-tier and new first-tier cities have forced many young people to turn their attention to the rental market, setting off an upsurge of housing rental. Based on the random forest model, this paper selects two cities, Shanghai and Wuhan, to study the price trend of the housing rental market and its influencing factors. Finally, it is found that the random forest regression model has no significant effect on the rental forecast in Shanghai. It may be that for a highly modernized first-tier city, the variables selected in this paper are not enough to fully explain the rental price. The prediction effect of rental price in Wuhan is significantly better, among which the characteristics of urban area and housing itself have a great impact on rental price. This research can serve as a reference for future researchers in the housing rental market, while helping landlords and tenants make optimal choices.

Keywords: *Rental price; Machine learning; Random forest algorithm*

1. INTRODUCTION

As a pillar industry to promote the development of the national economy, real estate is an indispensable part of the economic development of every country and region. In recent years, Chinese real estate has developed rapidly [1]. And the housing rental market has developed into a specialized market [2] from the stage of the welfare rental market [3] where the company provides preferential rooms to employees as welfare benefits. No region has a similar growth in such a long period of time in modern times. Therefore, it is undoubtedly worth exploring and learning for other developing countries.

With the gradual opening of the housing leasing market, the transaction scale of housing leasing has continued to expand and has become a significant part of our country's real estate market [4].In particular, the high housing prices in first-tier cities have added a lot of pressure to young people who are new to social work and ordinary people who have entered the city for work, making them turn to the rental market(Chinh Ho et al,2014) [5].According to the "Investigation Report on Rental Consumption Behavior" released by Anjuke in 2019 that the current trend of China's rental housing resources is concentrating in first-tier cities. With the increase of the number of people who rent houses, the proportion of people renting houses in first-tier cities and some new first-tier cities may exceed 40% in the future [6]. It can be seen that the rental market in first-tier cities has become a hot topic, and its future trend has attracted more and more attention.

Due to individual differences, everyone has diverse requirements, budgets and options for renting a house [7]. For the moment, the situation of information asymmetry between tenants and landlords in China's rental market is particularly prominent [8], which leads to an imbalance

G. Vilas Bhau et al. (Eds.): MSEA 2022, ACSR 101, pp. 587–593, 2023.
https://doi.org/10.2991/978-94-6463-042-8_84

between supply and demand. Therefore, building a feasible rental price prediction model through rental market data is a rigid demand in the current rental industry. Effective analysis of factors affecting rental prices and price forecasts can help tenants and landlords better understand the rental market and make optimal decisions [9].

Compared with foreign countries, there are few related researches on the price forecast of the rental market in our country. In recent years, the primary methods for predicting the rental market price in our country include multiple linear regression, neural network, and random forest.

With linear regression model, Z Wu found the main influenced factor including the number of rooms, income distribution, ratio of teachers and students [10]. Similarly, Linear regression model show the location is greatly related to the house price in Abdul Hafeed's research [11].

In contrast to multiple linear regression model, the artificial neural network models (ANN) seems to be more effective and accurate because of high goodness-of-fit values and low error values [12]-[13]. BEE-HUA GOH et al implied that compared to the multiple linear regression model, ANN has lower MAPE (mean absolute percentage error), hence it is more accurate [13]. Rahman et al also observe the MAPE as low as 4.41% or 4.55% in their research about the house price prediction based on ANN [14]. But they cannot find they cannot find exact factor that influence the house price because of the "black box" nature of ANN.

Recently, to avoid the "black box" nature of ANN [15], some scholars has applied random forest model (RF) to real estate price prediction as an alternative since it can capture the complexity or non-linearity of the market [16]. Hong et al. noted that "area" is the most important factor for price, followed by "number of buildings in the apartment complex" [16]; M Čeh et al. illustrated that, besides the high accuracy of prediction, the year of instruction and living area have significant impact on the price of apartment in Ljubljana [17].

In the past, scholars used random forest model to predict the price of first or second-hand commercial house, but their research are rarely focused on the rental house market of Chinese main city. This paper took the rental house price of Shang Hai and Wu Han for example, use random forest model predict their rental house price respectively, and compare their result. Our research will provide reference and conclusions for the closing section research.

The remainder of this paper is organized as follows:

Methodology introduces the process of data collection and exploration of data's feature, as well as the model development of random forest. The results and analysis of the research will be presented in Result and Discussion. Finally, the paper will be concluded in Conclusion.

2. METHODOLOGY

In this research, the random forest machine learning method is used to analyze the rent-affecting factors and forecast the housing rental prices in Shanghai and Wuhan. The research process mainly includes Data Collection, Data's feature which helps to understand the datasets and identify features in the dataset. Afterwards the model is developed using the proposed random forest algorithm.

2.1.Data's Collection

LIANJIA is a large real estate agency website with comprehensive data which reflect the current real market situation and which has been applied in many researches, for example, the research of repeat sale housing price by Xianling Yang [18].In the development of the model, this paper selects LIANJIA to collect data on the 2022's rental market of Shanghai and Wuhan which are regarded as the representatives of first-tier cities in China, and establishes the datasets of rental price and housing features of the two cities for research. In the datasets of rental house price and housing features in the two cities, there are 2546 items in Shanghai and 2377 items in Wuhan respectively, with the missing value which is less than 1%. Both datasets contain the rental house price and nine housing features of each house. As the random forest model is a supervised learning method based on datasets, it is necessary to divide the two urban datasets respectively: 80% of the data is set as training group, and 20% of the data is set as test group.

2.2.Data's Feature

Refer to the research on determinants of house prices by Jinlong Duan and Guangjin Tian [19], we select nine housing features in the datasets. The housing features variable presented in the datasets includes ARE which is the area in the cities, SUB which is whether near the subway, ELE which is whether has elevator equipment, DEC which is whether decorated in high quality, FA which is the floor area, ORI which is the orientation of house, the layout of house which is divided into three kinds of statistics in order to make the model data be quantified: BED which is the number of the bedrooms, LIV which is the number of living rooms and BAT which is the number of bathrooms. These features would be presented in the table as follows:

Table 1. Explanation of Variables

Variable	Explanation

ARE	Area in the cities
SUB	Whether near the subway
ELE	Whether has elevator equipment
DEC	Whether decorated in high quality
FA	Floor area
ORI	Orientation of house
BED	Number of the bedrooms
LIV	Number of living rooms
BAT	Number of bedrooms

2.3. Model Development

The random forest method used in this study is consistent with the research of Boston housing price prediction by Abigail Bola Adetunjia [20]. Random forest model is a method widely applied in the field of machine learning at present. It randomizes the use of variables and the use of data, generates multiple decision trees whose classification and regression are based on the repeated dichotomous data, and then summarizes the results of all decision trees to form a random forest. The random forest model improves the prediction accuracy without significantly increasing the computation amount, and it is insensitive to multiple linear variables. The results are robust to the unbalanced data and can well explain the effects of multiple explanatory variables. We use RStudio software to construct and test the model based on the datasets, and the model code can be searched in https://github.com/Hhutao/randon-forest-Rstudio. The research process of random forest algorithm adopted in this study is as follows:

1)Random forest modeling and data analysis of Shanghai rental price dataset:

a)Randomly select a specific proportion of data in the dataset to establish random data samples that can be applied by the program for learning and training of the random forest model.

b)Starting from the root node, the information gain of all possible features is calculated for the node, and the feature with maximum gain is selected as the node feature.

c)Establish child nodes with different values of features, recursively call the method in (2) for the child nodes, and build a decision tree until all feature information gain is small or there is no feature selection.

d)Repeat steps (1) to (3).

e)Summarize the training results of decision tree, combine them into a random forest, and select additional datasets for model testing.

2)Conduct random forest modeling and data analysis on Wuhan rental price dataset whose steps are the same as those of Shanghai dataset analysis.

3)Performance evaluation indicators were used to analyze the training and prediction effect of random forest model in Shanghai and Wuhan, and then feature importance were used to analyze the variables affecting rental prices in the two cities respectively and explore the way and degree of influence of the variables.

The Random forest model demonstration figure is as follow [21]:

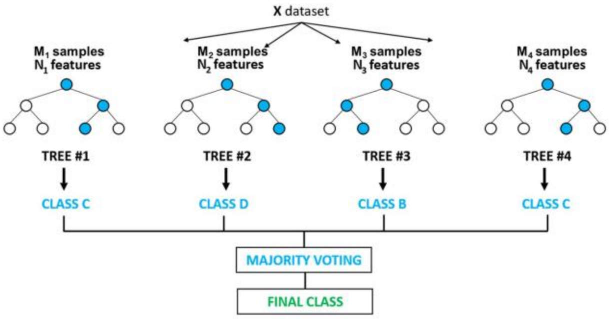
X dataset
M_1 samples N_1 features
M_2 samples N_2 features
M_3 samples N_3 features
M_4 samples N_4 features
TREE #1
TREE #2
TREE #3
TREE #4
CLASS C
CLASS D
CLASS B
CLASS C
MAJORITY VOTING
FINAL CLASS

Figure 1. Random forest model demonstration figure

3. Result and Discussion

3.1. Data Exploration

In order to better understand the information features of rental housing in Shanghai and Wuhan, descriptive statistics tables and correlation heat maps for four continuous variables of rental housing price and housing features will be displayed. The descriptive statistics table shows the value range and distribution of each variable, and the selected indicators include the total count of data, the data mean, the standard deviation, the minimum value, lower quartile, median, upper quartile and the maximum value. As the table shows, the rental price range in Shanghai is dramatically wide, with the highest reaching 70,000 RMB per month and the median also at a high level; In Shanghai, most houses are less than 100 square meters, and the smallest is 6.3 square meters. The layout of houses in Shanghai tends to be homogeneous, with most bedrooms, living rooms and bathrooms no more than 2. Rental price range in Wuhan is relatively small, and the price distribution is closer; Its floor area distribution is roughly the same as that of Shanghai, mainly within 100 square meters; However, Wuhan has a relatively diversity of housing layout, with three-bedroom houses accounting for at least 25% of the total number of rentals.

Table 2. Descriptive Statistics Table of Shanghai

	PRI	FA	BED	LIV	BAT
Count	2545.00	2545.000	2545.000	2545.000	2545.000
Mean	7521.852	68.995	1.717	1.167	1.132
Min	3400.00	6.300	1.000	0.000	0.000
25%	5500.00	42.380	1.000	1.000	1.000
50%	6200.00	59.030	2.000	1.000	1.000
75%	7800.00	89.490	2.000	2.000	1.000
Max	70000.000	502.560	5.000	3.000	5.000
Std	4539.143	40.139	0.792	0.592	0.415

Table 3.Descriptive Statistics Table of Wuhan

	PRI	FA	BED	LIV	BAT
Count	2376.000	2376.000	2376.000	2376.000	2376.000
Mean	2286.524	71.187	2.553	1.332	1.268
Min	650.000	5.100	1.000	0.000	0.000
25%	1300.000	30.000	2.000	1.000	1.000
50%	2300.000	79.2156	3.000	1.000	1.000
75%	2900.000	98.050	3.000	2.000	2.000
Max	12000.000	393.000	7.000	4.000	5.000
Std	1197.437	42.265	1.013	0.653	0.572

To obtain the correlation between various continuous variables, we used the correlation heat map. Correlation heat map is a common visualization method to show the degree of correlation between variables. Heat map represents data in the form of graphics and colors. The darker the color, the stronger the correlation between two variables. As shown in the figure, rental prices in Shanghai are strongly correlated with house area, reaching 0.78 which is a high degree, and they are also strongly correlated with the number of bedrooms and toilets. There is a strong correlation between rental price and housing area in Wuhan, reaching 0.84, but the correlation between rental price and housing layout is weak, and only the number of living rooms has a medium correlation of 0.55.

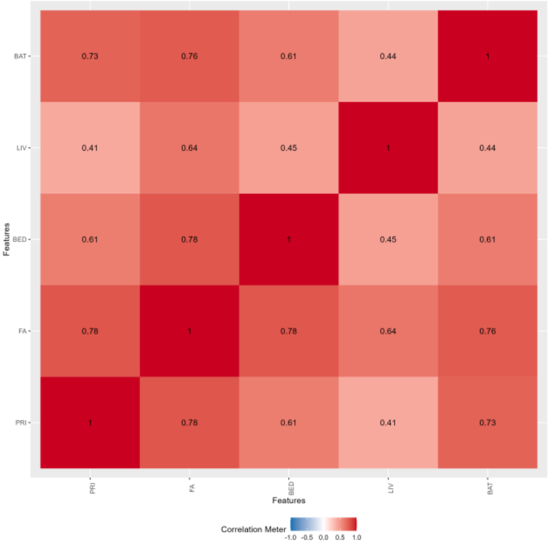

Figure 2. Correlation Heat Map of Shanghai

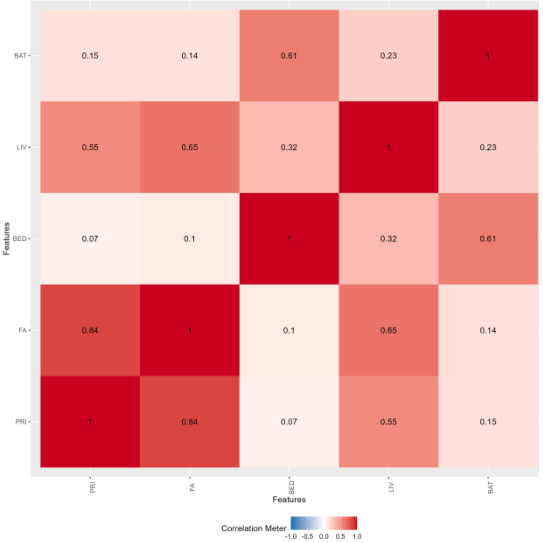

Figure 3. Correlation Heat Map of Wuhan

3.2. Model Test

After the development and training of the random forest model based on the datasets, the next stage is to test the predictive ability of the model obtained. The trained model is used to predict the rental price of the test group datasets, and then the rental price predicted by the model is compared with the actual rental prices of the test group data and the relative difference is calculated. In Shanghai and Wuhan, some rental price prediction results based on the test group data set are shown in the tables. Although the predicted value of rental house price cannot be exactly the same as the actual value, and some data even have a prediction deviation of 20%, the overall data is relatively fitting.

Table 4. Actual Vs Predicted Rent of Shanghai

S/N	Actual Rent	Predict Rent	Different
1	5200	4987.67	4.08%
2	5500	5484.07	0.29%
3	9000	7967.26	11.47%
4	8500	8984.38	5.69%
5	8300	7447.83	10.26%
6	8500	7403.47	12.91%
7	5500	5495.13	0.09%
8	5700	6153.78	7.96%
9	7200	6030.32	16.24%
10	6000	6209.89	3.49%
11	5000	5115.49	2.31%
12	9200	7205.26	21.68%
13	5500	6416.91	16.67%
14	6600	6598.58	0.02%
15	14600	13623.95	6.68%

Table 5. Actual Vs Predicted Rent of Wuhan

S/N	Actual Rent	Predict Rent	Different
1	3000	3134.49	4.48%

2	1300	1300	0.00%
3	3500	2792.81	20.20%
4	3000	3029.87	0.99%
5	2200	2585.63	17.52%
6	1000	1039.28	3.92%
7	3200	3641.02	13.78%
8	1150	1350.54	17.43%
9	1100	1135.43	3.22%
10	1800	2101.71	16.76%
11	1600	1755.69	9.73%
12	900	866.65	3.70%
13	950	1005.46	5.83%
14	3500	3947.74	12.79%
15	1700	1644.76	3.24%

3.3. Model Evaluation

After the model is established and tested, performance evaluation indicators are used to assess the performance of the model. These indicators are MAE, R2, and MRSE. After the model evaluation indicators are obtained, the study generates scatter plots to show the extent to which the predicted values deviate from the actual values. MAE is used to evaluate the average value of absolute error between parameter estimation value and parameter actual value in parameter estimation. RMSE is the square root of the ratio of the deviation between the predicted value and the actual value and the number of observations n. R-squared is the goodness of fit. As shown in the figure, RMSE and MAE of the test group of the rental price prediction model in Shanghai are 1779.80 and 922.16 respectively, indicating that there is a large deviation between the predicted value of rental price and the real value. However, the result of R-Squared is 0.8343, indicating that the variables selected by the model have a good explanation for the rental price in Shanghai. RMSE and MAE of the test group of rental price prediction model in Wuhan are 345.91 and 234.71 respectively, indicating that the deviation between the predicted value of rental price and the real value is relatively small, and the prediction effect of the model is good. R-squared result 0.9157 shows that the selected variables have a very good explanation effect on rental price prediction model of Wuhan.

Table 6. Evaluation of the Random Forest Model of Wuhan

Set	RMSE	R-squared	MAE
Train	1290.61	09283	652.82
Test	1779.80	08343	922.16

Table 7. Evaluation of the Random Forest Model of Wuhan

Set	RMSE	R-squared	MAE
Train	232.90	0.9628	152.31
Test	345.91	0.9157	234.71

Scatter plots are generated to show the extent to which the actual value deviates from the predicted value. The scatter plots show that the fitting degree between the predicted value and the real value is poor in Shanghai, while the fitting condition between the predicted value and the real value in Wuhan is good. The generated scatter diagram of predicted value and actual value of rental house price:

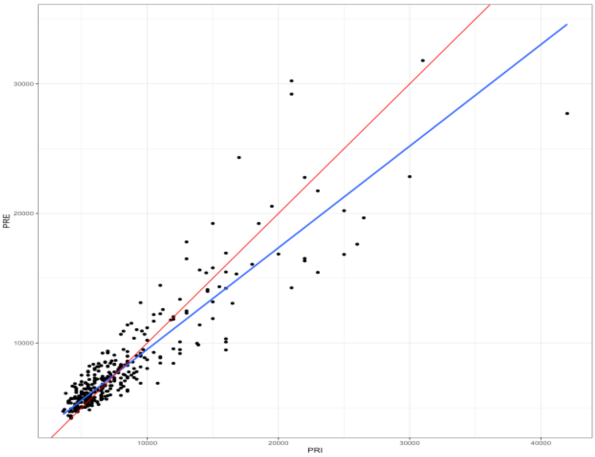

Figure 4. Scatter Plot Actual Vs Predicted of Shanghai

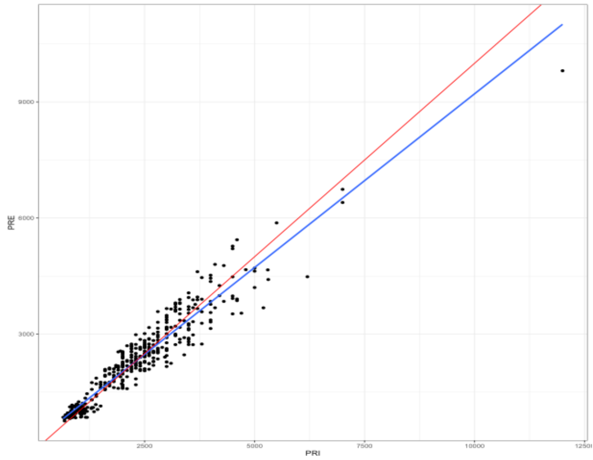

Figure 5. Scatter Plot Actual Vs Predicted of Wuhan

Finally, this paper will use feature importance to measure the way and degree of influence of each specific variable on rental price. As shown in the figure, for the rental market of Shanghai, the area in the cities has a great influence on the rental price. For example,the %IncMSE of Songjiang District and Huangpu District are both relatively high,with the digits of near 22.These two cities have a more significant influence on the rental price than other inner-city areas. In addition, the floor area , whether near the subway and whether has elevator equipment also play a relatively large impact on rental prices. Wuhan's rental market is more significantly affected by the area in the cities. The %IncMSE of Jianghan District and Wuchang District, two relatively developed areas in Wuhan, are both above 75.Furthermore,compared with the subway, the rental price in Wuhan is more affected by the floor area and whether has elevator equipment whose %IncMSE are 220.374 and 112.464 respectively.

Table 8. Feature Importance of Variables in Shanghai and Wuhan

Variable	%IncMSE	Variable	%IncMSE
ARE Changning	13.456	ARE Donghugaoxin	68.831
ARE Chongming	1.473	ARE Dongxihu	31.666
ARE Fengxian	1.178	ARE Hannan	1.592
ARE Hongkou	1.535	ARE Hanyang	13.909
ARE Huangpu	22.153	ARE Hongshan	12.350
ARE Jiadin	5.848	ARE Huangpi	16.400
ARE Jing'an	8.821	ARE Jiang'an	12.457
ARE Jinshan	2.603	ARE Jianghan	75.337
ARE Minxing	5.122	ARE Jiangxia	17.931
ARE Pudong	3.291	ARE Qiaokou	26.840
ARE Putuo	4.443	ARE Qingshan	3.951
ARE Qingpu	4.496	ARE Wuchang	79.823
ARE Songjiang	22.516	ARE Xinzhou	11.239
ARE Xuhui	11.120	ARE Zhuankou	17.669
ARE Yangpu	5.106	SUB	1.648
SUB	46.160	ELE	112.464
ELE	31.022	DEC	43.944
DEC	18.173	FA	220.374
FA	44.010	ORI_N	10.677
ORI_N	3.741	ORI_NE	1.200
ORI_NE	1.131	ORI_NW	4.724
ORI_NW	1.317	ORI_S	16.754
ORI_S	13.098	ORI_SE	6.663
ORI_SE	15.633	ORI_SW	2.657
ORI_SW	2.392	ORI_W	1.506
ORI_W	-0.135	BED	38.211
BED	21.864	LIV	24.129
LIV	21.103	BAT	41.778
BAT	17.704		

4. Conclusion

This paper applies the random forest model to forecast the rental house price in Wuhan and Shanghai and analyze the factors that affect the rental house price. There are 2 reasons why this paper pays attention to the rental markets in Wuhan and Shanghai. First, the development and unique performance of China's rental house market makes it a lot of potential research value. Secondly, in recent years, the rental house market in China's first-tier and new first-tier cities has attracted widespread attention in China, while Shanghai and Wuhan are representatives of China's first-tier cities and new first-tier cities, therefore their rental markets are representative. This research selected data from the LIANJIA website for research. The results show that the random forest model has a good prediction to the rental house market in Shanghai and Wuhan, and different influencing factors have different degrees of influence on the rental prices in the two cities.

This study finds that the random forest model has a poor prediction effect on rental prices in Shanghai. Although the selected variables have a good explanatory effect on rental prices in Shanghai, the average difference between the predicted value and the real value is large. Researchers assume that the variables selected in this research are not enough to fully explain the rental price in Shanghai. For first-tier cities as Shanghai, the rental price may be affected by more factors. Therefore, the next research of analyzing the rental market of Shanghai in the future will consider more possible influencing factors. For Wuhan, the random forest model has significantly better prediction effect of rental house price, and the fitting degree between the predicted price and the real price is high, indicating that the variables selected in this paper can explain the rental house price in Wuhan well. The analysis of feature importance shows that area in the city, such as Jianghan District and Wuchang District, have a significant impact on rental prices in Wuhan. In addition, the feature of housing itself, such as floor area and elevator equipment, has a high impact on rental prices. This study uses the random forest method to study the rental market in Wuhan and Shanghai, which provides reference for rental market researchers in the future and helps renters and landlords make optimal decisions in the rental market.

References

[1] Ahuja, A., Cheung, L., Han, G., Porter, N. and Zhang, W, “Are house prices rising too fast in China?” IMF Working Paper WP/10/274. Retrieved from http://www.imf.org/external/pubs/ft/wp/2010/wp10274.pdf

[2] Eftimoski M, McLoughlin K, “Housing policy and economic growth in China Australia. Australia: Reserve Bank of Australia,”2019.

[3] X. YANG,” The Experience of Housing Welfare Policy in Developed Countries and Its Enlightenment on China —Based on the Perspective of the Structural reform at Supply-side,” West Forum, vol.27, pp.107-115, January 2014.

[4] W. Yuan,” Analysis on the current situation of long-term rental housing market expedited by China's policy dividend and Discussion on the establishment of double custody mechanism,” Contemporary Economics, pp.7-9, 2020.

[5] Chinh Ho, David Hensher,” Housing Prices and Price Endogeneity in Tenure and Dwelling Type Choice Models,” Case Studies on Transport Policy, pp.107-115, 2014.

[6] Anjuke,” Investigation report on rental consumption behavior,”2019.3.15, https://www.donews.com/news/detail/4/3039153.html.

[7] PRESTON V, TAYLOR S M," Personal Construct Theory and Residential Choice," Annals of the Association of American Geographers, pp.437-451, 1981.

[8] L. Yang, K. W. Chau, Y. Chen," Impacts of information asymmetry and policy shock on rental and vacancy dynamics in retail property markets," Habitat International,2021.

[9] P.H. Hendershott," Bubbles in Metropolitan Housing Markets," pp.191–207, 1996.

[10] Z. Wu," Prediction of California House Price Based on Multiple Linear Regression," Academic Journal of Engineering and Technology Science, pp.11-15,2020

[11] A. Abdulhafedh," Incorporating Multiple Linear Regression in Predicting the House Prices Using a Big Real Estate Dataset with 80 Independent Variables," Open Access Library Journal, January 2022.

[12] H. Selim," Determinants of house prices in Turkey: Hedonic regression versus artificial neural network," Expert systems with Applications, 2009.

[13] B.GOH," Forecasting Residential Construction Demand in Singapore: A Comparative Study of the Accuracy of Time Series, Regression and Artificial Neural Network Techniques," Engineering, Architecture and Construction Management, pp.261–275, 1998.

[14] S. Rahman, N. Maimun, M. Razali, S. Ismail," The artificial neural network model (ANN) for Malaysian housing market analysis," Journal of the Malaysian Institute of Planners, vol. 17, pp.1–9, January 2019.

[15] J. Tu," Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes," Journal of clinical epidemiology, vol. 49, pp.1225–1231, November 1996.

[16] J. Hong, H. Choi, W. Kim," A house price valuation based on the random forest approach: the mass appraisal of residential property in South Korea," International Journal of Strategic Property Management, vol. 24, pp.140–152, 2020.

[17] M. Čeh, M. Kilibarda, A. Lisec, B. Bajat," Estimating the performance of random forest versus multiple regression for predicting prices of the apartments," ISPRS international journal of Geo-information, July 2018.

[18] X. Yang, H. Xu," The Impact of Brokers on the Formation Mechanism of Repeat Sale Housing Price: An Empirical Analysis Based on the Data of Beijing HOMELINK," Journal of Central University of finance, vol. 9, pp.82–93, January 2018.

[19] J. Duan, G Tian, L Yang, T Zhou," Addressing the macroeconomic and hedonic determinants of housing prices in Beijing Metropolitan Area, China," Habitat International,2021.

[20] A. Adetunjia, O. Noah Akande, F. Ajala, O. Oyewo, Y. Akande, G. Oluwadara, "House Price Prediction using Random Forest Machine Learning Technique," Science Direct,2022.

[21] Zhihu," The Ensemble Model:Random Forest Model(2),"2018.6.25, https://zhuanlan.zhihu.com/p/38484624.