Second-Hand Housing Market Research Based on Random Forest Algorithm

Zhiyuan Bai
Jinhe Center for Economic Research, Xi'an Jiaotong University, Xi'an 710049, China.

Abstract: The research on second-hand houses mainly faces the following problem. what are the factors that affect the price of second-hand houses? Focusing on the problem of second-hand houses, based on data mining and data analysis, and with the help of software such as Stata, this paper establishes the random forest model, and explains the factors that affect the price of second-hand houses. Through the prediction of the random forest model, this paper concludes that the factors affecting the price of second-hand houses are ranked as follows: the number of primary and secondary schools > the age of the house > the building area > the number of supermarkets > the number of scenic spots > the number of subways > the degree of decoration > the number of bedrooms, and gives some suggestions to people buying second-hand hourses.

Keywords: Second-Hand Housing; Data Analysis; Random Forest Model; Prediction

Introduction

On the evening of May 28, Xi'an (China) Housing and Urban-rural Development Bureau released the new 528 house purchase policy, which included changes in house purchase qualifications, identification of just-needed houses, and changes in provident fund down payment interest rates. Subsequently, the price trend of second-hand housing in Xi'an continued to deviate from the trend of new housing prices. After a month-on-month decrease of 0.1% in April, the downward trend continued in May, and the decline was even wider, with a month-on-month decrease of 0.4%. From the point of view of products, the prices of second-hand houses of $90m^2$ and below, 90 to $144m^2$ and above $144m^2$ in Xi'an in May all dropped month-on-month by 0.1%, 0.5% and 0.5% respectively. Therefore, in order to be able to provide reasonable advice to people buying second-hand homes, an effective model needs to be established to explain this change.

In order to study the second-hand housing market in China, this paper selects Xi'an as a reference city, because it is located in the central part of China and belongs to a high-speed development city, which is very representative.

In order to explore various factors affecting the price of second-hand housing in Xi'an, and establish a mathematical model. In this paper, the obtained data is processed, and the qualitative and quantitative variables are assigned according to certain rules through the assignment method. In addition, the influencing factors are also screened, and the information that can affect the second-hand housing prices is screened out by one-way ANOVA analysis of variance. A support vector machine regression model and a random forest model are established to explore various factors that affect the price of second-hand housing in Xi'an.

In order to test the rationality of the model and improve its accuracy, this paper uses the crawler pair to crawl the data of other months to test the model. The model is improved from three aspects: reducing the caliber difference, using multiple months of data for forecasting, and removing future data to improve the accuracy of the model.

1. Literature Review

The study of regional housing prices belongs to the scope of urban computing, and more and more researches have started to focus on urban computing in recent years, and many scholars have found many latent laws in cities through urban big data, which can help city builders to make decisions. For example, machine learning can be used to solve the problems of area function identification and the placement of shared bicycles [1], and to predict human mobility by exploring POI data and geographic information data [2], etc. All kinds of studies show that there are extensive correlations among urban big data, and these correlations provide information clues for the continuous intelligence of cities.

As the main economic pillar of today's economy and society, real estate has a profound impact on the nation, whether it is used for immediate housing needs or for financial investment. For home buyers, housing prices are an important factor influencing their decisions [3]. The exploration of the trend of house price changes from different perspectives has always been a hot issue of concern. For the problem of house price prediction, Yuming Gao and Renjin Zhang [4] improved BP neural network by genetic algorithm and obtained high prediction accuracy in house price prediction in Guiyang city, but did not qualitatively analyze the house price in conjunction with local policies.

2. Research methods

2.1 Data processing

The characteristic indicators are quantified with the following principles:

(a) Building area (X_1)

According to the Code for Urban Residential Design, a house below $90m^2$ is a small house, 90-$144m^2$ is a medium house, and a house above $144\,m^2$ is a large house. According to this method, this paper divide the floor area of second-hand houses into three categories: large, small, medium and small, and make the value of large house type 2, medium house type 1 and small house type 0 respectively.

(b) House age(X_2)

Building age using direct quantification method, setting house age

$$X_2 = 2022 - \text{building age}$$

Where building age is the year that the house was built.

(c) Number of living rooms (X_3)

Home buyers have the greatest preference for 3 living roooms, and the utility brought by the degree of decoration (X_4) decreases in the order of finishing, simple decoration, and rough, so this paper use the Likert scale method to rate them, as shown in the following table:

Table 1: Data processing rating scale for the number of living rooms

Number of living rooms	Scores
3	3
2 and 4	2
1 and 5	1
0 and 6	0

Table 2: Data processing rating scale for the degree of decoration

Degree of decoration	Scores
finishing	2
simple	1
rough	0

(d) House orientation (X_5)

After consulting the information on the property website, this paper got people's preference of orientation as follows: due south > southeast > southwest > east > northeast > north or west > northwest, so the same Likert scale method was used to score.

(e) Floor(X_6)

According to the literature study, people prefer second homes with elevators. Among them, Article 6 of the Residential Design Code stipulates that buildings with 7 or more floors must be installed with elevators, so this paper specify that residences with a number of floors greater than or equal to 7 are high-rise residences and those with less than 7 floors are non-high-rise residences.

(f) Number of elementary school (X_7), number of subways (X_8), number of supermarkets (X_9), and number of attractions (X_{10})

Based on the location information of each neighborhood, this paper searched in Baidu map and investigated the number of primary and secondary schools within 1km radius (X_7), the number of subways within 600m radius (X_8) and the number of supermarkets within 400m radius (X_9) of each neighborhood, considering that the number of famous monuments in Xi'an is large and the number of monuments may also have some influence on the house price, this paper also counted the number of attractions within a 600m radius of each neighborhood (X_{10}). Finally, this paper used the direct quantification method to assign values to each element by their numbers.

2.2 Screening of indicators

After processing the above indicators, it is necessary to determine whether they have a significant effect on house prices. The method this paper use here is one-way ANOVA. The p-values of the variables adjusted by Bartlett's test are shown in the following table:

Table 3: The p-values of each variable

X_1	X_2	X_3	X_4	X_5
0.000	0.000	0.000	0.073	0.000
X_6	X_7	X_8	X_9	X_{10}
0.060	0.000	0.001	0.000	0.001

From the above analysis, it is clear that the variables are all significant at a significance level of 10%, indicating that they are all significant influences on house prices.

2.3 Model Building

2.3.1 Building a random forest model

Random forest regression is the process of generating many decision trees by randomly sampling the sample observations and feature variables of the modeled dataset separately. Each sampling result is one tree, and each tree generates rules and judgment values that match its own attributes. The forest eventually integrates the rules and judgment values of all decision trees.

2.4 Result

The following are the results of the random forest model:

Table 4: Random forest Regression Results

Training set R2 score	Training set MSE	Validation set R2 score	Validation set MSE
0.89	575968	0.54	3488190

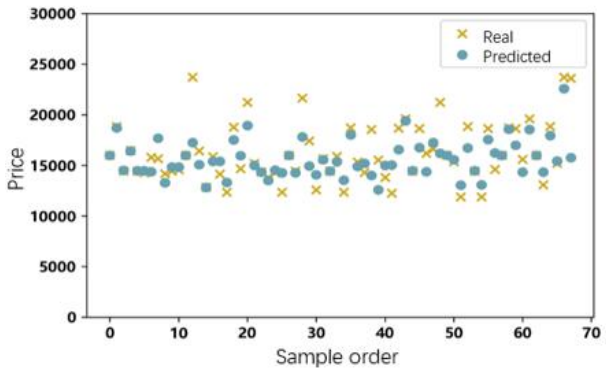

Figure 1: Training set fitting results of random forest model

It can be seen that the random forest model has a better degree of adaptation to this question, so this paper choose the random forest model as the training method for this question. Also, this paper obtained the characteristic importance of the respective variables, where the number of primary and secondary schools > age of the house > floor area > number of supermarkets > number of attractions > number of subways > decoration > living room.

2.5 Model Testing

This paper uses the second-hand housing price data in June to establish a model for the influencing factors of housing prices, and now looks for the data of second-hand housing in Xi'an in May to test the rationality of the model obtained in the first question and make improvements.

2.6 Test for plausibility

This paper call the Baidu map API to query the surrounding information of the corresponding place, and obtain the data in the same format as the first question through the same data processing method.

This paper directly use the random forest model obtained by fitting the housing price data in June to analyze the obtained May data. The data analysis is as follows:

Table 5: Random forest model results table for May data

R2 score	MSE
-0.54	277943210

2.7 Model improvements

In order to reduce the caliber difference, this paper used the June data crawled on Fangtianxia website (https://xian.esf.fang.com/) to fit our model, and then predicted the May data after fitting.The results are as follows:

Table 6: Random forest model results after reducing the caliber difference

R2 score	MSE
0.445	99738328

It can be seen that the fitting effect has been improved.

2.8 Improvement 2: Use multiple months of data for forecasting

Considering that data from other months has not been utilized, this paper use all months except May to make predictions, and the results are as follows:

Table 7: Random forest model results using Multi-Month Forecasts

R2 score	MSE
0.567	77678016

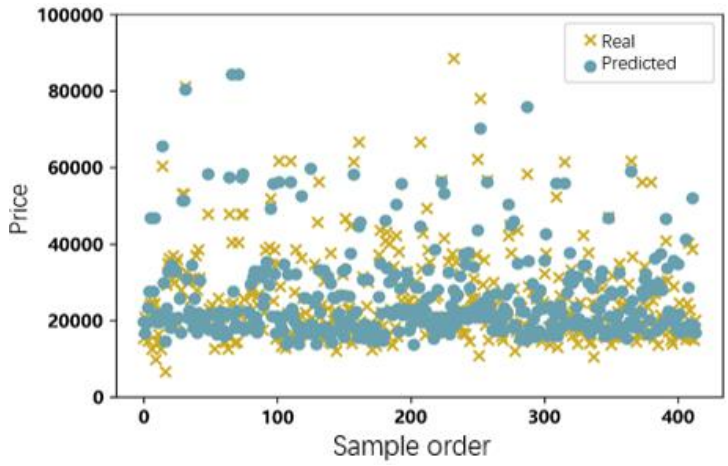

Figure 2: The fitting results of using the random forest model training set for multi-month forecasts

It can be seen that the accuracy of the model has been greatly improved: the value of R^2 is further increased to 0.567, and the value of MSE is further reduced by about 22.12%.

3. Discussion

3.1 Conclusions

This paper obtains the feature importance ranking of the number of primary and secondary schools > the age of the house > the building area > the number of supermarkets > the number of attractions > the number of subways > the decoration > the bedroom.

Based on the random forest model obtained by fitting, the obtained May data was analyzed to test the rationality of the obtained model, and R^2 =-0.54, MSE=277943210. After analyzing the reasons for its poor fitting effect, this paper first improves the first improvement by reducing the caliber difference, and obtains R^2=0.445, MSE=99738328, and then uses the data of multiple months to make the second improvement, the accuracy of the model has been greatly improved.

References

[1] Chen, L., Zhang, D., Pan, G., et al. (2015, September). Bike sharing station placement leveraging heterogeneous urban open data. In Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (pp. 571-575).

[2] Wang, M., Yang, S., Sun, Y., & Gao, J. (2016, December). Predicting human mobility from region functions. In 2016 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) (pp. 540-547). IEEE.

[3] Soltani, A., Pettit, CJ., Heydari, M., & Aghaei, F. (2021). Housing price variations using spatio-temporal data mining techniques. Journal of Housing and the Built Environment, 36(3), 1199-1227.

[4] Gao, Y., & Zhang, R. (2014). Analysis of house price prediction based on genetic algorithm and BP neural network. Comput. Eng, 40(4), 187-191.