Artificial Intelligence Evolution
https://ojs.wiserpub.com/index.php/AIE/

UNIVERSAL WISER
PUBLISHER

Research Article

Data Analysis on the Influencing Factors of the Real Estate Price

Huang-Mei He[1], **Yi Chen**[1], **Jia-Ying Xiao**[1], **Xue-Qing Chen**[1], **Zne-Jung Lee**[2*]

[1]School of Big Data, Fuzhou University of International Studies and Trade, Fuzhou, China
[2]School of Intelligent Construction, Fuzhou University of International Studies and Trade, Fuzhou, China
Email: lrz@fzfu.edu.cn

Received: 26 May 2021; **Revised:** 17 July 2021; **Accepted:** 2 September 2021

Abstract: China has carried out a large number of real estate market reforms that change the real estate market demand considerably. At the same time, the real estate price has soared in some cities and has surpassed the spending power of many ordinary people. As the real estate price has received widespread attention from society, it is important to understand what factors affect the real estate price. Therefore, we propose a data analysis method for finding out the influencing factors of real estate prices. The method performs data cleaning and conversion on the used data first. To discretize the real estate price, we use the *mean* ± *standard deviatio*n (*SD*), *mean* ± 0.5 *SD*, and *mean* ± 2 *SD* of the price and divide it into three categories as the output variable. Then, we establish the decision tree and random forest model for six different situations for comparison. When the data set is divided into training data (70%) and testing data (30%), it has the highest testing accuracy. In addition, by observing the importance of each input variable, it is found that the main influencing factors of real estate price are cost, interior decoration, location, and status. The results suggest that both the real estate industry and buyers should pay attention to these factors to adjust or purchase real estate.

***Keywords*:** data analysis, real estate price, decision tree, random forest

1. Introduction

The continuous soaring of real estate prices has not only increased the life pressure of citizens but also caused the premature overdraft of social consumption [1-3]. While demanding considerable bank loans, the price increase promotes the rents for commercial office buildings and factories which impact the economy. This has made the real estate industry become one of the major industries in the development of the national economy, and the fluctuation of real estate prices is affected by many factors. As the fluctuation is related to the various interests of people, they begin to analyze real estate prices. Substantial changes in the prices bring about many problems. It not only affects the national economy and peoples' lives but also other industries [4]. Real estate prices are related to the self-interests of people as well as the fiscal revenues of the government at all levels and the stability of the national financial system.

Recently, machine learning and data mining algorithms such as decision tree regression, neural networks, support regression vector, fuzzy regression model, and multiple linear regression have been applied to predict real estate prices [5-10]. Proposed approaches have admirable techniques for prediction. However, it also needs to understand what factors affect real estate prices for implementing the regulation policy of real estate prices [11, 12]. Therefore, we propose a method by focusing on the influencing factors instead of the regression model for real estate prices. The result

DOI: https://doi.org/10.37256/aie.222021966

is expected to contribute to the sustainable, stable, and healthy development of the real estate market.

This article is organized as follows. The second part mainly describes the methodology of data pre-processing, including raw data processing, the cleaning of missing values, and the conversion of data units. The third part explains how to divide the data into six sets to establish decision trees and random forest models for predicting real estate prices based on the model and how to test the importance of the influencing factors of real estate prices. The fourth part presents the discussions. Finally, based on the above discussions, the last part concludes the research result by proposing the influencing factors of real estate prices.

2. Data preprocessing

The data for this research is provided by Fujian Jianke Real Estate Appraisal. The data set has a total of 100 instances. The original data must undergo data conversion and cleaning to meet the needs of the data analysis. Missing values in the data set are included as variables. The data is converted for the unit price of 10,000 yuan/m^2 and the transaction price of 10,000 yuan. To establish the data analysis model, the output variable Y.price in the data set is discretized into three variables by calculating the mean and standard deviation.

2.1 *The classification with mean ± SD*

The *mean* ± *SD* is used to divide the value area of the classification. Values less than *mean* − *SD* are grouped in category 1 (low). Those in the range of between *mean* ± *SD* are grouped in category 2 (medium) while those greater than *mean* + *SD* are category 3 (high). The *mean* value and *SD* of the Y.price are 2.445647 and 0.9165614, respectively. The value of the *mean* ± *SD* is 1.529085 to 3.362208. Figure 1 shows that the three categories (low, medium, and high) have 15, 72, and 13 observations, respectively.

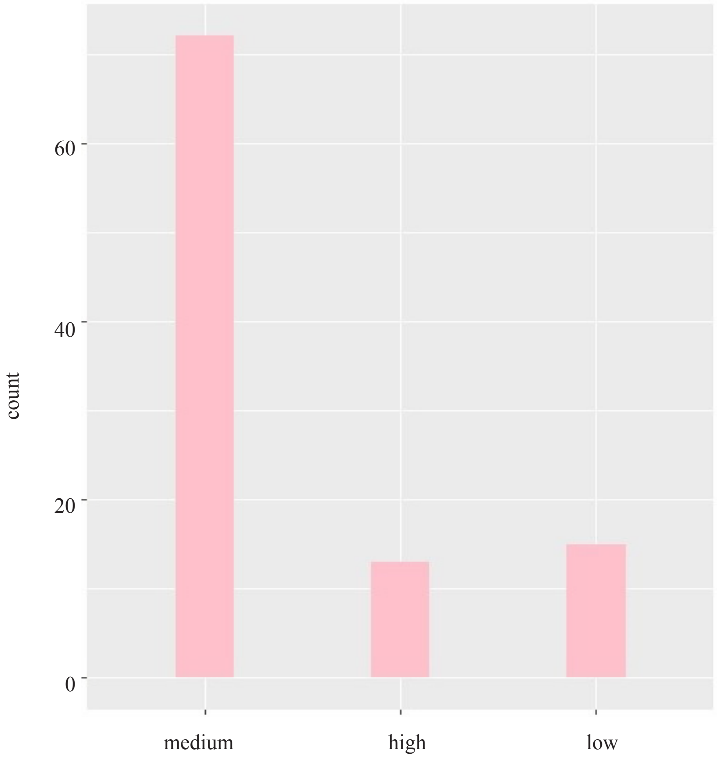


Figure 1. The classification with *mean* ± *SD*

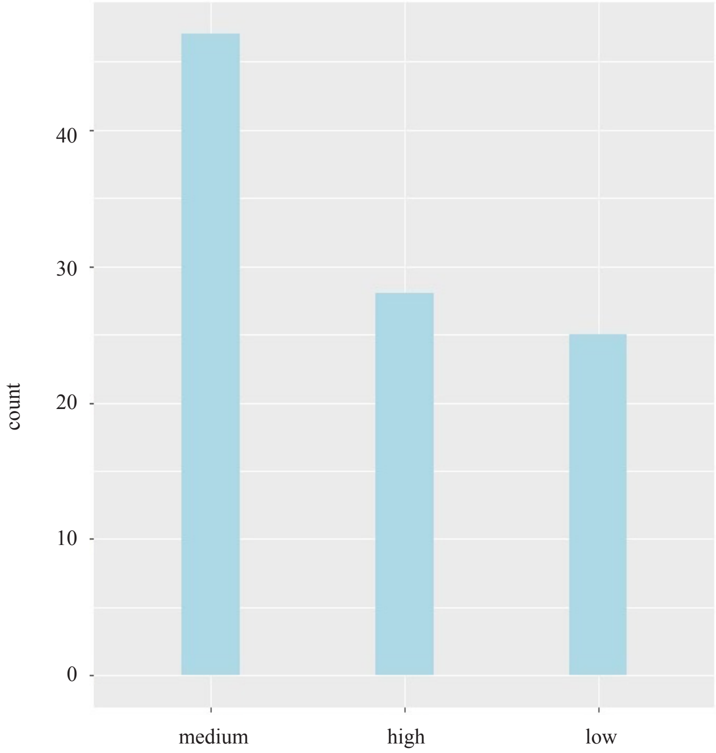


Figure 2. The classification with *mean* ± 0.5 *SD*

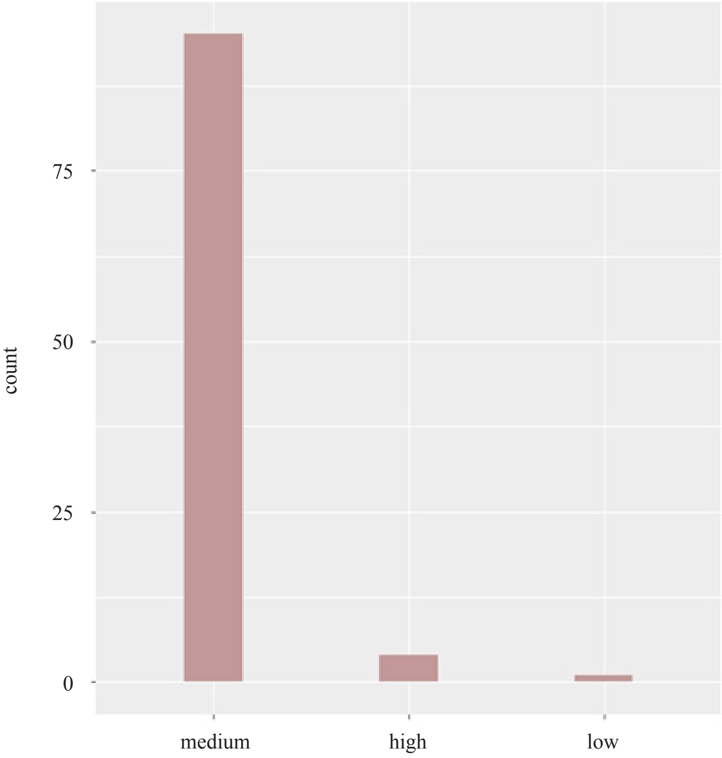


Figure 3. The classification with *mean* ± 2 *SD*

2.2 *The classification with mean ± 0.5 SD*

Values less than *mean* − 0.5 *SD* are grouped into category 1 (low), those between *mean* ± 0.5 *SD* into category 2 (medium), and those greater than *mean* + 0.5 *SD* into category 3 (high). The value of the *mean* ± 0.5 *SD* is 1.987366 to 2.903927. Figure 2 shows that there are 25, 47, and 28 observations in category 1 (low), 2 (medium), and 3 (high).

2.3 *The classification with mean ± 2 SD*

Values less than *mean* − 2 *SD* are grouped into category 1 (low), those between *mean* ± 2 *SD* into category 2 (medium), and those greater than *mean + 2 SD* into category 3 (high). The value of the *mean* ± 2 *SD* is 0.6125239 to 4.27877. Figure 3 shows that categories 1 (low), 2 (medium), and 3 (high) include 1, 95, and 4 observations, respectively.

3. Introduction to decision tree and random forest model

In machine learning, decision trees are models for classification and regression. They can predict and classify data and know what proportions of variables are the most important. As a method of knowledge representation, the decision tree algorithm is generally used for classification tasks. In the classification problem, the decision tree is the step of classifying observations based on each attribute. Generally speaking, a decision tree (Figure 4) is composed of a root node, several inner nodes, and several leaf nodes [13, 14]. The formation of a decision tree includes two main stages: the spanning tree stage and the decision tree pruning stage. Before starting to build the decision tree, the data set can be divided into training and testing data sets. The training data set is for model construction, and the testing data set is used to truly evaluate the performance of the model.

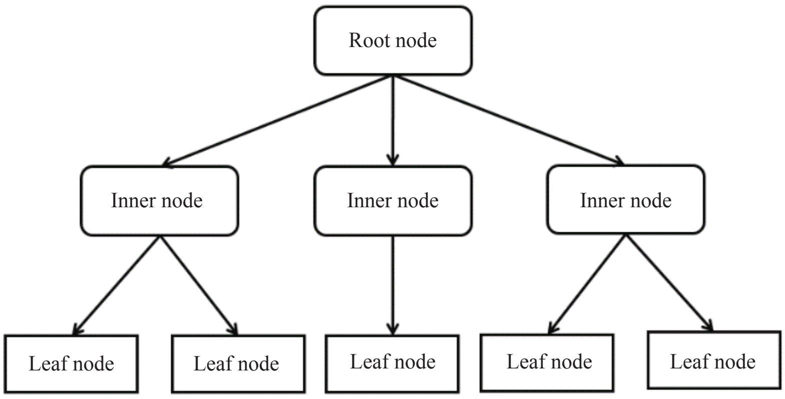


Figure 4. Tree structure diagram of decision tree

The random forest method is a combination of tree predictors [15, 16]. It is inspired by the decision tree algorithm and improves the decision tree algorithm to form a new type of algorithm. Random forest is different from the formation process of the classification and regression algorithm, which is manifested in the following aspects. The method of generating the training data set is different. Random forest requires the number of training data to be consistent with the number of original data sets. The splitting rules of the tree are different. Each node must be based on the principle of minimum node impurity. The pruning operation is not performed in the random forest. The impurity measurement method is the *Gini index*. If a dataset D with m data points is split into k subsets $\{D_1, D_2, ..., D_k\}$ with sizes $\{m_1, m_2, ..., m_k\}$ respectively, the *Gini index* is represented as follows.

$$Gini_{split}(D)=\sum_{l=1}^{k}\frac{m_l}{m}Gini(D_l) \tag{1}$$

When dataset D is divided into D_1 and D_2 according to the point σ.

$$D_1=\{(x,y)\in D\,|\,D(x)=\sigma\},\ D_2=D-D_1 \tag{2}$$

The *Gini index* is defined as follows.

$$Gini(D,\sigma)=\frac{m_1}{m}Gini(D_1)+\frac{m_2}{m}Gini(D_2) \tag{3}$$

3. The used model and results

3.1 *Decision tree model*

The data used in this article is to set the ratio of the training data set and the testing data set to two cases: 70% training data and 30% testing data; 80% training data and 20% testing data. The decision trees are shown from Figure 5 to Figure 10. Since the classification of the data set is divided into three cases and the proportion of the data set is set to two cases, the decision tree model has the following forms.

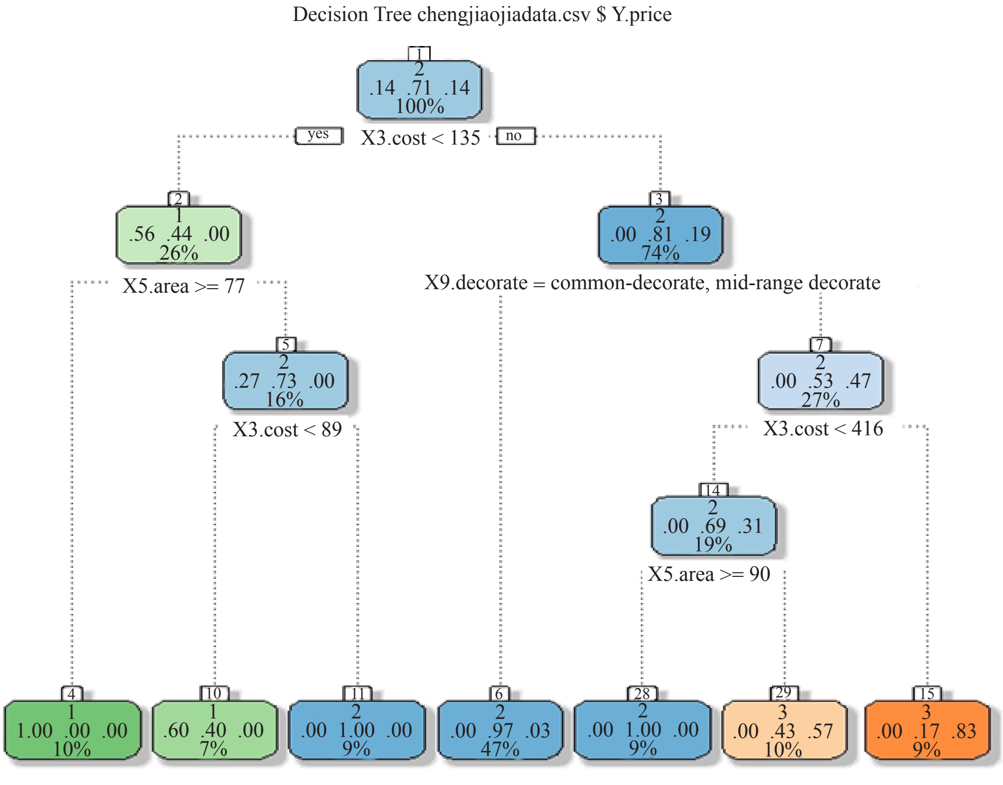


Figure 5. Decision tree model of 70% training data and 30% testing data with *mean* ± *SD*

(1) The ratio of the data set is 70% training data and 30% testing data with *mean* ± *SD*. From Figure 5, it can be

seen that there are mainly seven rules for the branch structure of the decision tree model.

Rule 1: If the transaction price is less than 1.35 million yuan and the area is greater than or equal to 77 square meters, the unit price level is 1 (low) accounting for 10%.

Rule 2: If the transaction price is less than 1.35 million yuan and the area is less than 77 square meters, and the transaction price is greater than or equal to 890,000 yuan, the unit price level is 1 (low) accounting for 7%.

Rule 3: If the transaction price is less than 1.35 million yuan and the area is less than 77 square meters, and the transaction price is less than 890,000 yuan, the unit price level is 2 (medium) accounting for 9%.

Rule 4: If the transaction price is greater than or equal to 1.35 million yuan and the decoration is ordinary decoration and mid-range decoration, the unit price level is 2 (medium) accounting for 47%.

Rule 5: If the transaction price is greater than or equal to 1.35 million yuan and the decoration is not ordinary or mid-range decoration, and the transaction price is less than 4.16 million yuan, and the area is greater than or equal to 90 square meters, the unit price level is 2 (medium) accounting for 9%.

Rule 6: If the transaction price is greater than or equal to 1.35 million yuan and the decoration is not ordinary or mid-range decoration, and the transaction price is less than 4.16 million yuan, and the area is less than 90 square meters, the unit price level is 3 (high) accounting for 10%.

Rule 7: If the transaction price is greater than or equal to 1.35 million yuan and the decoration is not ordinary and mid-range decoration, and the transaction price is greater than or equal to 4.16 million yuan, the unit price level is 3 (high) accounting for 9%.

(2) The ratio of the data set is 70% training data and 30% testing data with *mean* ± 0.5 *SD*. From Figure 6, the decision tree model with *mean* ± 0.5 *SD* also has seven branch structure rules.

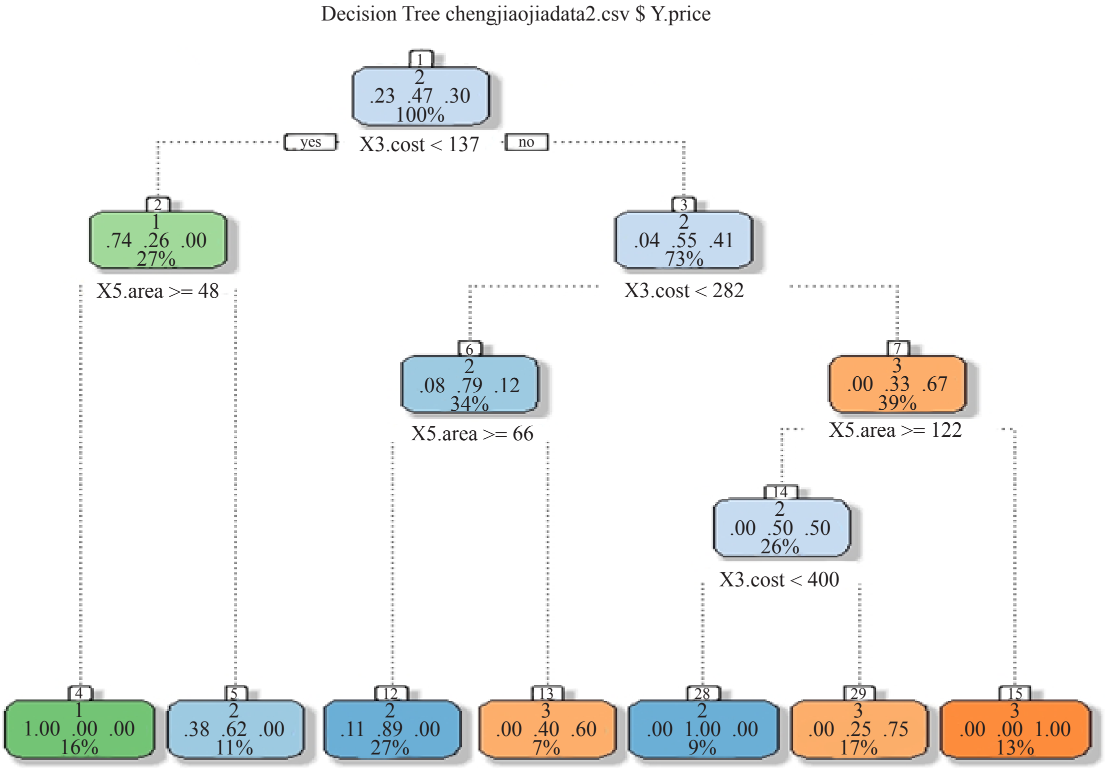


Figure 6. Decision tree model of 70% training data and 30% testing data with *mean* ± 0.5 *SD*

Rule 1: If the transaction price is less than 1.37 million yuan and the area is greater than or equal to 48 square meters, the unit price level is 1 (low).

Rule 2: If the transaction price is less than 1.37 million yuan and the area is less than 48 square meters, the unit price level is 2 (medium).

Rule 3: If the transaction price is greater than or equal to 1.37 million yuan and the transaction price is less than 2.82 million yuan, and the area is greater than or equal to 66 square meters, the unit price level is 2 (medium).

Rule 4: If the transaction price is greater than or equal to 1.37 million yuan and the transaction price is less than 2.82 million yuan, and the area is less than 66 square meters, the unit price level is 3 (high).

Rule 5: If the transaction price is greater than or equal to 1.37 million yuan and the transaction price is greater than or equal to 2.82 million yuan, and the area is greater than or equal to 122 square meters, and the transaction price is less than 4 million yuan, the unit price level is 2 (medium).

Rule 6: If the transaction price is greater than or equal to 1.37 million yuan and the transaction price is greater than or equal to 2.82 million yuan, and the area is greater than or equal to 122 square meters, and the transaction price is greater than or equal to 4 million yuan, the unit price level is 3 (high).

Rule 7: If the transaction price is greater than or equal to 1.37 million yuan and the transaction price is greater than or equal to 2.82 million yuan, and the area is less than 122 square meters, the unit price level is 3 (high).

(3) The ratio of the data set is 70% training data and 30% testing data with *mean* ± 2 *SD*. From Figure 7, it can be seen that there are five rules for the branch structure of the decision tree model.

Rule 1: If the decoration is ordinary decoration, high-end decoration, mid-range decoration, the unit price level is 2 (medium).

Rule 2: If the decoration is not ordinary decoration, high-end decoration, mid-range decoration, and the area is less than 39 square meters, the unit price level is 1 (low).

Rule 3: If the decoration is not ordinary decoration, high-end decoration, mid-range decoration, and the area is greater than or equal to 39 square meters, and the transaction price is less than 3.52 million yuan, the unit price level is 2 (medium).

Rule 4: If the decoration is not ordinary decoration, high-end decoration, mid-range decoration, and the area is greater than or equal to 39 square meters, and the transaction price is greater than or equal to 3.52 million yuan, and the floor is greater than or equal to the 8th floor, the unit price level is 2 (medium).

Rule 5: If the decoration is not ordinary decoration, high-end decoration, mid-range decoration, and the area is greater than or equal to 39 square meters, and the transaction price is greater than or equal to 3.52 million yuan, and the floor is less than the 8th floor, the unit price level is 3 (high).

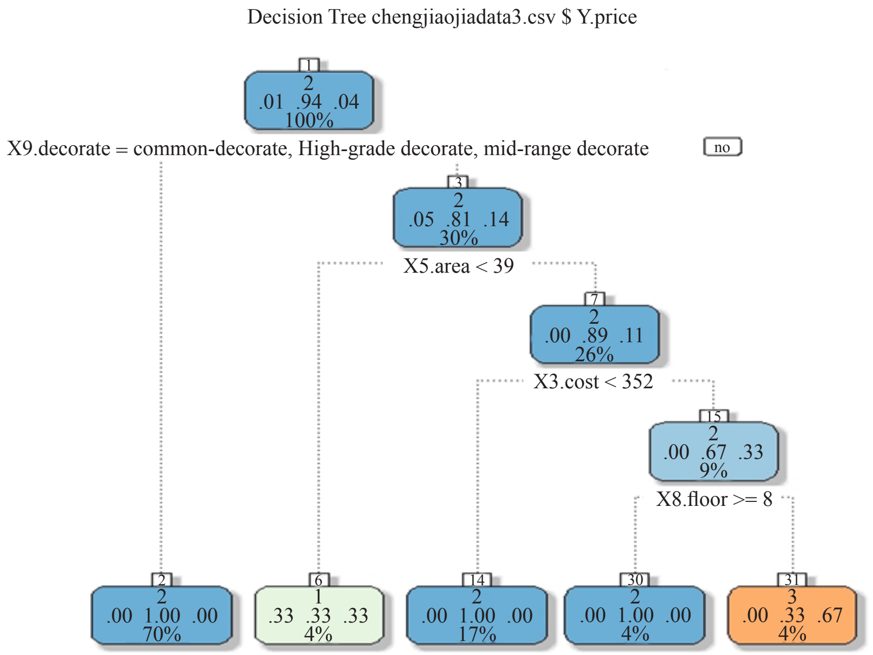


Figure 7. Decision tree model of 70% training data and 30% testing data with *mean* ± 2 *SD*

(4) The ratio of the data set is 80% training data and 20% testing data with *mean* ± *SD*. From Figure 8, it can be seen that there are six rules for the branch structure of the decision tree model.

Rule 1: If the transaction price is less than 1.35 million yuan and the area is greater than or equal to 78 square meters, the unit price level is 1 (low).

Rule 2: If the transaction price is less than 1.35 million yuan and the area is less than 78 square meters, and the transaction price is less than 860,000 yuan, the unit price level is 1 (low).

Rule 3: If the transaction price is less than 1.35 million yuan and the area is less than 78 square meters, and the transaction price is greater than or equal to 860,000 yuan, the unit price level is 2 (medium).

Rule 4: If the transaction price is greater than or equal to 1.35 million yuan and the decoration is ordinary decoration, mid-range decoration, or simple decoration, the unit price level is 2 (medium).

Rule 5: If the transaction price is greater than or equal to 1.35 million yuan and the decoration is not ordinary, mid-range, or simple, and the transaction price is less than 2.8 million yuan, the unit price level is 2 (medium).

Rule 6: If the transaction price is greater than or equal to 1.35 million yuan and the decoration is not ordinary, mid-range, or simple, and the transaction price is greater than or equal to 2.8 million yuan, the unit price level is 3 (high).

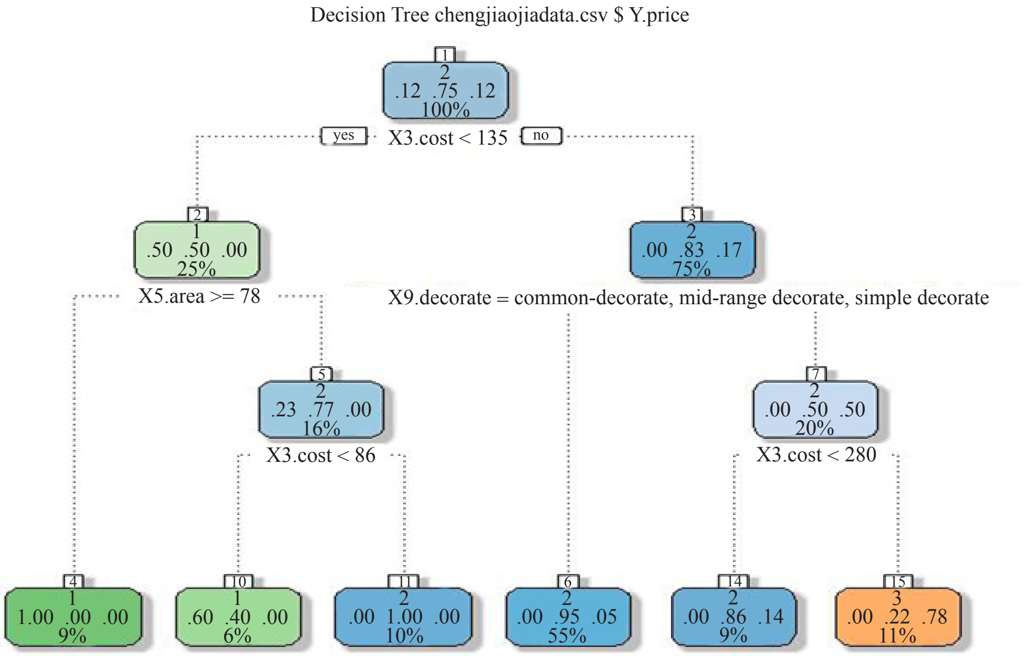


Figure 8. Decision tree model of 80% training data and 20 testing data with *mean* ± *SD*

(5) The ratio of the data set is 80% training data and 20% testing data with *mean* ± 0.5 *SD*. From Figure 9, it can be seen that there are eight rules for the branch structure of the decision tree model.

Rule 1: If the transaction price is less than 1.37 million yuan and the area is greater than or equal to 48 square meters, the unit price level is 1 (low).

Rule 2: If the transaction price is less than 1.37 million yuan and the area is less than 48 square meters, the unit price level is 2 (medium).

Rule 3: If the transaction price is greater than or equal to 1.37 million yuan and the transaction price is less than 2.62 million yuan, and the area is greater than or equal to 66 square meters, the unit price level is 2 (medium).

Rule 4: If the transaction price is greater than or equal to 1.37 million yuan and the transaction price is less than 2.62 million yuan, and the area is less than 66 square meters, the unit price level is 3 (high).

Rule 5: If the transaction price is greater than or equal to 1.37 million yuan and the transaction price is greater than or equal to 262, and the decoration is ordinary decoration, simple decoration, rough, and the transaction price is less than 4.74 million yuan, and the area is greater than or equal to 117 square meters, the unit price level is 2 (medium).

Rule 6: If the transaction price is greater than or equal to 1.37 million yuan and the transaction price is greater than or equal to 262, and the decoration is ordinary decoration, simple decoration, rough, and the transaction price is less than 4.74 million yuan, and the area is less than 117 square meters, the unit price level is 3 (high).

Rule 7: If the transaction price is greater than or equal to 1.37 million yuan and the transaction price is greater than or equal to 262, and the decoration is ordinary decoration, simple decoration, or rough, and the transaction price is greater than or equal to 4.74 million yuan, the unit price level is 3 (high).

Rule 8: If the transaction price is greater than or equal to 1.37 million yuan and the transaction price is greater than or equal to 262, and the decoration is not ordinary, simple, or rough, the unit price level is 3 (high).

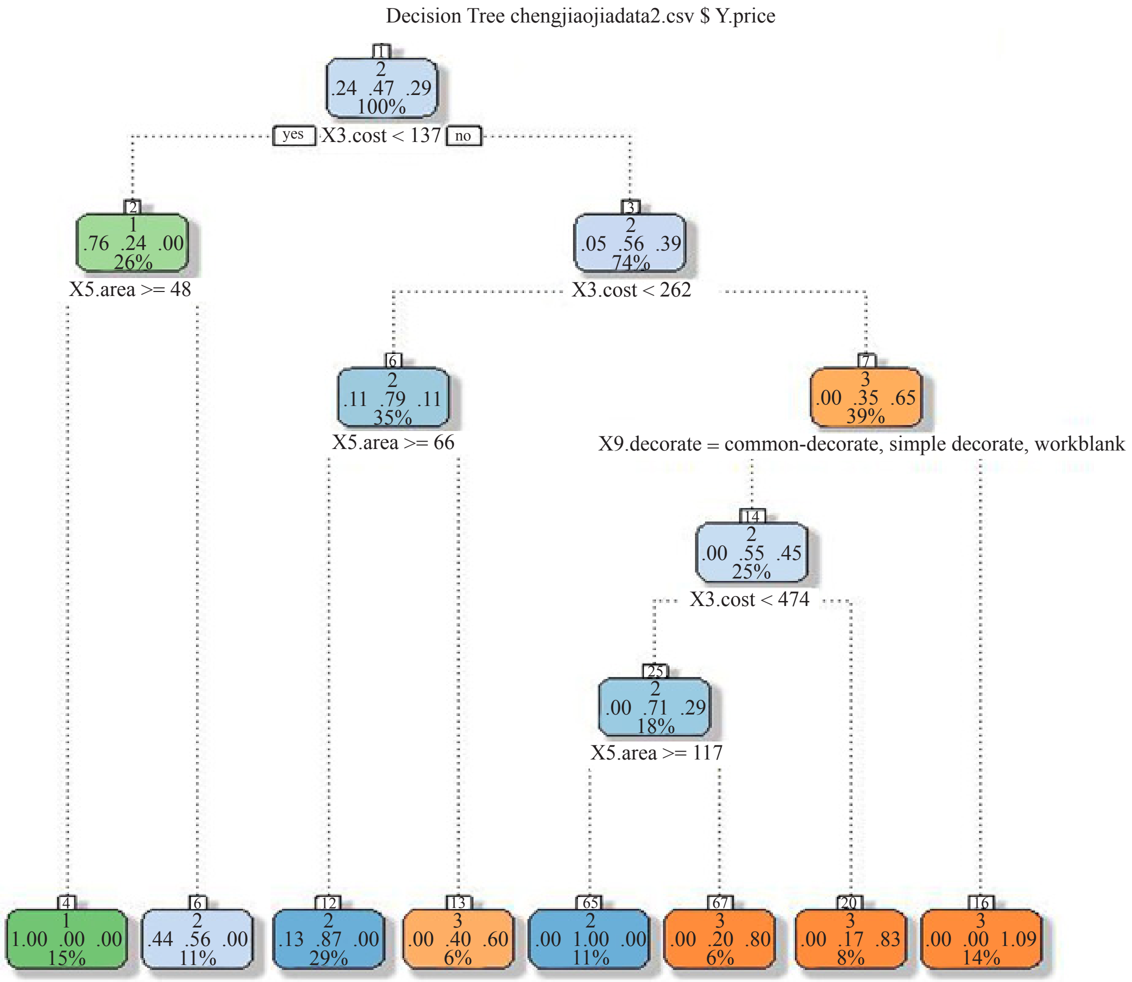


Figure 9. Decision tree model of 80% training data and 20% testing data with *mean* ± 0.5 *SD*

(6) The ratio of the data set is 80% training data and 20% testing data with *mean* ± 2 *SD*. From Figure 10, it can be seen that there are five rules for the branch structure of the decision tree model.

Rule 1: If the decoration is ordinary decoration, mid-range decoration, high-end decoration, the unit price level is 2 (medium).

Rule 2: If the decoration is not ordinary decoration, mid-range decoration, high-end decoration, and the area is less than 39 square meters, the unit price level is 1 (low).

Rule 3: If the decoration is not ordinary decoration, mid-range decoration, high-end decoration, and the area is greater than or equal to 39 square meters, and the transaction price is less than 3.52 million yuan, the unit price level is 2 (medium).

Rule 4: If the decoration is not ordinary decoration, mid-range decoration, high-end decoration, and the area is greater than or equal to 39 square meters, and the transaction price is greater than or equal to 3.52 million yuan, and the floor is greater than or equal to the 8th floor, the unit price level is 2 (medium).

Rule 5: If the decoration is not ordinary decoration, mid-range decoration, high-end decoration, and the area is greater than or equal to 39 square meters, and the transaction price is greater than or equal to 3.52 million yuan, and the floor is less than the 8th floor, the unit price level is 3 (high).

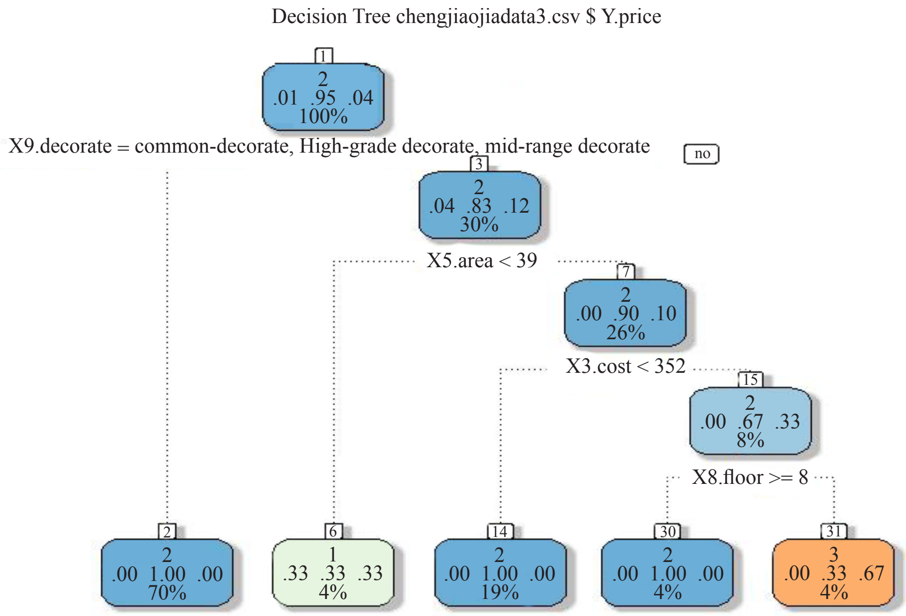
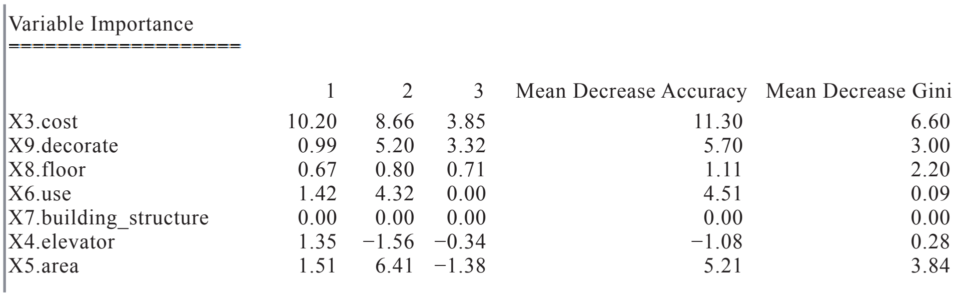


Figure 10. Decision tree model of 80% training data and 20% testing data with *mean* ± 2 *SD*

3.2 *Random forest model*

Random forest construction is set to randomly select two variables at each node of each tree to generate 100 decision trees. The results are shown from Figure 11 to Figure 16. Similar to the decision tree model, the random forest model is also divided into six forms for comparison.

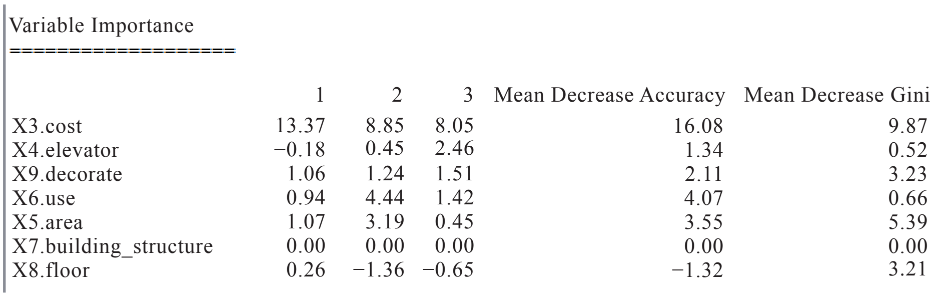


```
Variable Importance
===================

                          1      2      3   Mean Decrease Accuracy  Mean Decrease Gini
X3.cost               10.20   8.66   3.85                    11.30                6.60
X9.decorate            0.99   5.20   3.32                     5.70                3.00
X8.floor               0.67   0.80   0.71                     1.11                2.20
X6.use                 1.42   4.32   0.00                     4.51                0.09
X7.building_structure  0.00   0.00   0.00                     0.00                0.00
X4.elevator            1.35  −1.56  −0.34                    −1.08                0.28
X5.area                1.51   6.41  −1.38                     5.21                3.84
```

Figure 11. The variable importance of random forest of 70% training data and 30% testing data with *mean* ± *SD*

(1) The ratio of the data set is 70% training data and 30% testing data with *mean* ± *SD*. Figure 11 shows the result. In Figure 11, the “Mean Decrease Accuracy” and “Mean Decrease Gini” are two important indicators of the random forest model. Among them, the first one represents the decrease in the accuracy of the random forest prediction. The larger the value, the more important the variable. The second one represents the effect of each variable on the heterogeneity of the observed value at each node in the classification tree. The larger the value, the more important the variable. From Figure 11, we can know that the “Mean Decrease Accuracy” value and “Mean Decrease Gini” value of X3.cost is the largest, followed by X9.decorate, X5.area, and X6.use, indicating that these four input variables are relative to several other variables.

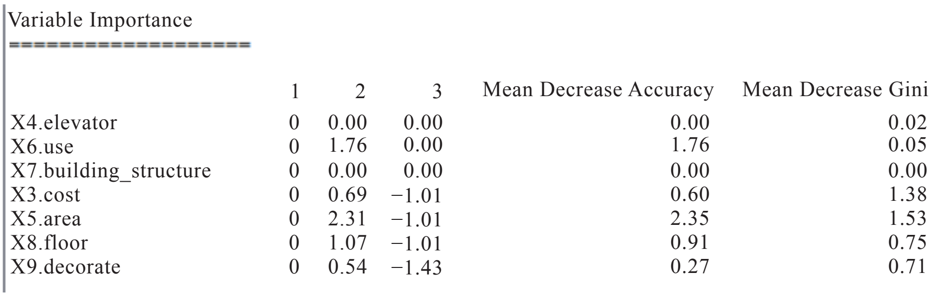
(2) The ratio of the data set is 70% training data and 30% testing data with *mean* ± 0.5 *SD*. From the Accuracy value and Gini value, it can be seen that the four variables of X3.cost, X6.use, X9.decorate, and X5.area are still relatively large. They are very important in the influencing factors of real estate prices. (Figure 12).



```
Variable Importance
===================

                           1      2      3   Mean Decrease Accuracy   Mean Decrease Gini
X3.cost                13.37   8.85   8.05                    16.08                 9.87
X4.elevator            −0.18   0.45   2.46                     1.34                 0.52
X9.decorate             1.06   1.24   1.51                     2.11                 3.23
X6.use                  0.94   4.44   1.42                     4.07                 0.66
X5.area                 1.07   3.19   0.45                     3.55                 5.39
X7.building_structure   0.00   0.00   0.00                     0.00                 0.00
X8.floor                0.26  −1.36  −0.65                    −1.32                 3.21
```

Figure 12. The variable importance of random forest of 70% training data and 30% testing data with *mean* ± 0.5 *SD*

(3) The ratio of the data set is 70% training data and 30% testing data with *mean* ± 2 *SD*. The largest accuracy and Gini values in Figure 13 are X3.cost, X6.use, and X5.area. These three variables are very important.



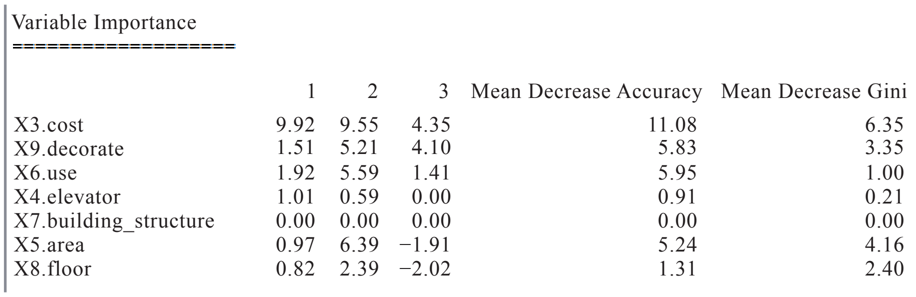
```
Variable Importance
===================

                       1     2      3   Mean Decrease Accuracy   Mean Decrease Gini
X4.elevator            0  0.00   0.00                     0.00                 0.02
X6.use                 0  1.76   0.00                     1.76                 0.05
X7.building_structure  0  0.00   0.00                     0.00                 0.00
X3.cost                0  0.69  −1.01                     0.60                 1.38
X5.area                0  2.31  −1.01                     2.35                 1.53
X8.floor               0  1.07  −1.01                     0.91                 0.75
X9.decorate            0  0.54  −1.43                     0.27                 0.71
```

Figure 13. The variable importance of random forest of 70% training data and 30% testing data with *mean* ± 2 *SD*

(4) The ratio of the data set is 80% training data and 20% testing data with *mean* ± 0.5 *SD*. Using the accuracy value and the Gini value to compare the importance, we can know that the variables X3.cost, X6.use, X9.decorate, and X5.area are more important among the factors affecting real estate price. (Figure 14).

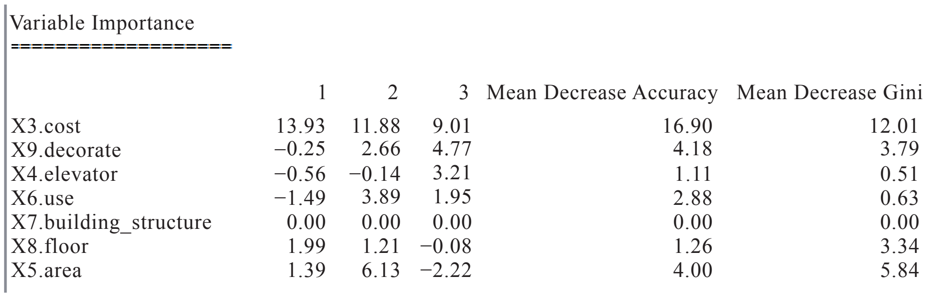
(5) The ratio of the data set is 80% training data and 20% testing data with *mean* ± *SD*. It can be seen from the figure that the variables X3.cost, X6.use, X9.decorate, and X5.area are more important. (Figure 15).

(6) The ratio of the data set is 80% training data and 20% testing data with *mean* ± 2 *SD*. It can be seen that these four variables X3.cost, X6.use, X9.decorate, and X5.area are still more important. (Figure 16).



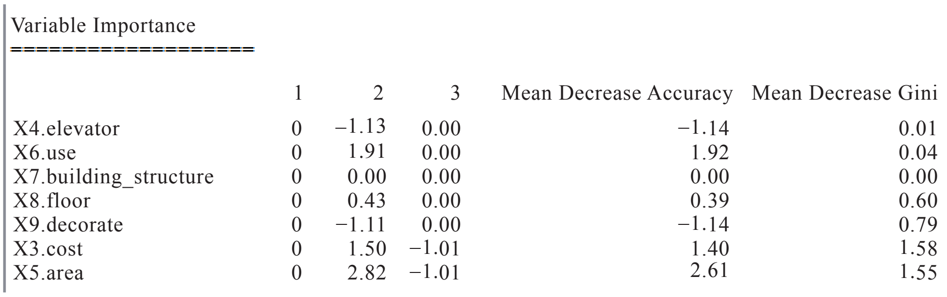
```
Variable Importance
===================

                          1     2      3  Mean Decrease Accuracy  Mean Decrease Gini

X3.cost                9.92  9.55   4.35                   11.08                6.35
X9.decorate            1.51  5.21   4.10                    5.83                3.35
X6.use                 1.92  5.59   1.41                    5.95                1.00
X4.elevator            1.01  0.59   0.00                    0.91                0.21
X7.building_structure  0.00  0.00   0.00                    0.00                0.00
X5.area                0.97  6.39  −1.91                    5.24                4.16
X8.floor               0.82  2.39  −2.02                    1.31                2.40
```

Figure 14. The variable importance of random forest of 80% training data and 20% testing data with *mean* ± *SD*



```
Variable Importance
===================

                           1      2      3  Mean Decrease Accuracy  Mean Decrease Gini

X3.cost                13.93  11.88   9.01                   16.90               12.01
X9.decorate            −0.25   2.66   4.77                    4.18                3.79
X4.elevator            −0.56  −0.14   3.21                    1.11                0.51
X6.use                 −1.49   3.89   1.95                    2.88                0.63
X7.building_structure   0.00   0.00   0.00                    0.00                0.00
X8.floor                1.99   1.21  −0.08                    1.26                3.34
X5.area                 1.39   6.13  −2.22                    4.00                5.84
```

Figure 15. The variable importance of random forest of 80% training data and 20% testing data with *mean* ± 0.5 *SD*



```
Variable Importance
===================

                        1      2      3  Mean Decrease Accuracy  Mean Decrease Gini

X4.elevator             0  −1.13   0.00                   −1.14                0.01
X6.use                  0   1.91   0.00                    1.92                0.04
X7.building_structure   0   0.00   0.00                    0.00                0.00
X8.floor                0   0.43   0.00                    0.39                0.60
X9.decorate             0  −1.11   0.00                   −1.14                0.79
X3.cost                 0   1.50  −1.01                    1.40                1.58
X5.area                 0   2.82  −1.01                    2.61                1.55
```

Figure 16. The variable importance of random forest of 80% training data and 20% testing data with *mean* ± 2 *SD*

4. Discussions

The data used in this article is to set the ratio of the training data set and the testing data set to two cases: 70% training data and 30% testing data; 80% training data and 20% testing data. Since the classification of the data set is divided into three cases and the proportion of the data set is set to two cases, the decision tree model has the following forms.

(1) The ratio of the data set is 70% training data and 30% testing data with *mean* ± *SD*. The testing accuracy of the model is 93.7%.

(2) The ratio of the data set is 70% training data and 30% testing data with *mean* ± 0.5 *SD*. The testing accuracy of the decision tree model is 84.6%.

(3) The ratio of the data set is 70% training data and 30% testing data with *mean* ± 2 *SD*. The testing accuracy of this model is 87.9%.

(4) The ratio of the data set is 80% training data and 20% testing data with *mean* ± *SD*. The testing accuracy of the decision tree model is 87.8%.

(5) The ratio of the data set is 80% training data and 20% testing data with *mean* ± 0.5 *SD*. The testing accuracy of

the decision tree model is 84.3%.

(6) The ratio of the data set is 80% training data and 20% testing data with *mean* ± 2 *SD*. The testing accuracy rate of the decision tree model is 88.1%.

Random forest construction is set to randomly select two variables at each node of each tree to generate 100 decision trees. Similar to the decision tree model, the random forest model is also divided into six forms for comparison.

(1) The ratio of the data set is 70% training data and 30% testing data with *mean* ± *SD*. The testing accuracy of the random forest model is 96.7%.

(2) The ratio of the data set is 70% training data and 30% testing data with *mean* ± 0.5 *SD*. The testing accuracy of random forest is 98%.

(3) The ratio of the data set is 70% training data and 30% testing data with *mean* ± 2 *SD*. The testing accuracy of random forest is 88.9%.

(4) The ratio of the data set is 80% training data and 20% testing data with *mean* ± *SD*. The testing accuracy of random forest is 93.4%.

(5) The ratio of the data set is 80% training data and 20% testing data with *mean* ± 0.5 *SD*. The testing accuracy of the random forest model is 96.5%.

(6) The ratio of the data set is 80% training data and 20% testing data with *mean* ± 2 *SD*. The testing accuracy of the random forest model is 88.9%.

The results of testing accuracy for the decision tree and random forest model are listed in Table 1 and Table 2. Judging from the decision tree and random forest model established in the above six forms, the first form (that is, the data set ratio is 70% training data and 30% testing data) has the best testing accuracy. For the decision tree, the three input variables X3.cost, X9.decorate (interior decoration), and X5.area (location) are the most important influencing factors. According to these three input variables, the price of the output variable Y.price can be predicted and classified. According to the variable importance given by the random forest model, we can divide the importance of influencing factors into: X3.cost, X6.use (status), X9.decorate, X5.area.

Table 1. The testing accuracy for the decision tree model

The data set ratio is 70% training data and 30% testing data		The data set ratio is 80% training data and 20% testing data	
mean ± *SD*	93.7%	*mean* ± *SD*	87.8%
mean ± 0.5 *SD*	84.6%	*mean* ± 0.5 *SD*	84.3%
mean ± 2 *SD*	87.9%	*mean* ± 2 *SD*	88.1%

Table 2. The testing accuracy for the random forest model

The data set ratio is 70% training data and 30% testing data		The data set ratio is 80% training data and 20% testing data	
mean ± *SD*	96.7%	*mean* ± *SD*	93.4%
mean ± 0.5 *SD*	98%	*mean* ± 0.5 *SD*	96.5%
mean ± 2 *SD*	88.9%	*mean* ± 2 *SD*	88.9%

5. Conclusions and future works

In this paper, we propose data analysis on the influencing factors of the real estate price. The used dataset is obtained from the company of Fujian Jianke Real Estate Appraisal. First, the decision tree model is used to find the decision rules for decision-makers. Thereafter, we can know that the best accuracy obtained from the data set is divided into 70% training data and 30% testing data by using the decision tree and random forest model. Moreover, the best accuracy of the decision tree model with *mean* ± *SD* is 93.7% and the random forest with *mean* ± 0.5 *SD* is 98%. The higher the score of mean decrease accuracy or Gini, the higher the importance of the factor in the model of real estate prices. From the decision tree and random forest model, the main influencing factors of real estate price are cost (X3. cost), interior decoration (X9.decorate), location (X5.area), and status (X6.use). Therefore, buyers should pay more attention to these factors and use these factors to predict the real estate price. In addition, the real estate industry can also adjust prices appropriately through these influencing factors. In the future work, authors will use other approaches such as weighted concept lattice and Hamming distance, Shannon entropy, and deep learning to do more research for finding out the influencing factors of real estate price [17-19].

Acknowledgement

This paper was support by Fuzhou University of International Studies and Trade, Grant No. FWXXS19111 and 2018KYTD-02.

References

[1] Yang H, Zhang HT, Zhao Y. Structural evolution of real estate industry in China: 2002-2017. *Structural Change and Economic Dynamics*. 2021; 57: 45-56.

[2] Li YC. The impact of the epidemic on China's real estate industry in 2020. In *6th International Conference on Financial Innovation and Economic Development (ICFIED 2021)*. Atlantis Press; 2021. p. 217-221.

[3] Fonseka M, Tian GL, Al Farooque O. Impact of environmental information disclosure and real estate segments on cost of debt: Evidence from the Chinese real estate industry. *Economics of Transition and Institutional Change*. 2020; 28(1): 195-221.

[4] Wang B. The evolving real estate market structure in China. *Understanding China's Real Estate Markets.* 2021; 9-19.

[5] Rawool AG, Rogye DV, Ranr SG. *House Price Prediction Using Machine Learning*. 2021.

[6] Li DY, Xu W, Zhao H, Chen RQ. A SVR based forecasting approach for real estate price prediction. In *2009 International Conference on Machine Learning and Cybernetics.* IEEE; 2009. p. 970-974.

[7] Sarip AG, Hafez MB, Daud MN. Application of fuzzy regression model for real estate price prediction. *Malaysian Journal of Computer Science*. 2016; 29(1): 15-27.

[8] Salem H, Mazzara M. ML-based Telegram bot for real estate price prediction. *Journal of Physics: Conference Series*. 2020; 1694(1): 012010.

[9] Uzut ÖG, Buyrukoglu S. Prediction of real estate prices with data mining algorithms. *Euroasia Journal of Mathematics Engineering Natural and Medical Sciences*. 2020; 7(9): 77-84.

[10] Wu ZX. Prediction of California house price based on multiple linear regression. *Academic Journal of Engineering and Technology Science*. 2020; 3(7): 11-15.

[11] Barreca A, Curto R, Rolando D. Urban vibrancy: an emerging factor that spatially influences the real estate market. *Sustainability*. 2020; 12(1): 346.

[12] Singla HK, Samanta PK. Identification of critical success factors (CSFs) for real estate developers (REDs) in India. *Journal of Financial Management of Property and Construction*. 2021. Available from: doi: 10.1108/JFMPC-04-2020-0028.

[13] Ghiasi MM, Zendehboudi S, Mohsenipour AA. Decision tree-based diagnosis of coronary artery disease: CART model. *Computer methods and programs in biomedicine*. 2020; 192: 105400.

[14] Charbuty B, Abdulazeez A. Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*. 2021; 2(01): 20-28.

[15] Chen YY, Zheng WZ, Li WB, Huang YM. Large group activity security risk assessment and risk early warning based on random forest algorithm. *Pattern Recognition Letters*. 2021; 144: 1-5.

[16] Lee ZJ, Lee CY, Yuan XJ, Chu KC. Rainfall forecasting of landslides using support vector regression. In *2020 3rd IEEE International Conference on Knowledge Innovation and Invention (ICKII)*. IEEE; 2020. p. 1-3.

[17] Singh PK. Cloud data processing using granular based weighted concept lattice and Hamming distance. *Computing*.

2018; 100(10): 1109-1132.

[18] Singh PK, Cherukuri AK, Li J. Concepts reduction in formal concept analysis with fuzzy setting using Shannon entropy. *International Journal of Machine Learning and Cybernetics*. 2017; 8(1): 179-189.

[19] Lee ZJ, Yang ZY, Lee CY, Chen ZH, Wu WB. Using improved neural network for the risk assessment of information security. In *IOP Conference Series: Materials Science and Engineering*. IOP Publishing; 2021. Available from: doi: 10.1088/1757-899X/1113/1/012025.