Frontiers in Business, Economics and Management
ISSN: 2766-824X | Vol. 10, No. 2, 2023

A House Price Prediction Model Based on K-means Clustering and Random Forest in Guangzhou

Zhishang Huang[1], Guanren Lai[2,*]

[1]School of Statistics and Mathematics, Guangdong University of Finance and Economics, Guangzhou, China
[2]School of Data Science, Guangzhou Huashang College, Guangzhou, China
[*]Corresponding author: 1448339928@qq.com

Abstract: This paper addresses the key issues in house price forecasting from multiple perspectives by establishing a house price forecasting model for Guangzhou city, providing valuable information and decision support for home buyers, developers, and the government. First, this paper employs the Person coefficient, stepwise regression model and t-test to address the problem of quantifying data and exploring house price factors. By analyzing the correlation between the relevant variables and house prices, the key characteristics that have significant and strong correlation effects on house prices are obtained. Second, the K-means clustering model was used to classify houses into three categories: economic houses, comfortable houses and high-end houses. This classification result provides a more detailed data base for the subsequent construction of the house price prediction model. Finally, the random forest house price prediction model was established in this paper, and the model was validated by error analysis and stability analysis. The average absolute value error and goodness-of-fit obtained were 0.08 and 0.92, respectively, indicating that the model has high accuracy and reliability. The research in this paper has important implications for all parties, including home buyers, developers, and the government. For home buyers, the model can help them better understand the market situation; for developers, the model can guide their reasonable pricing and development strategies; for the government, the model can provide a scientific basis for real estate market regulation and policy control to promote market stability and sustainable development.

Keywords: Person Correlation Coefficient, Stepwise Regression Model, Elbow Principle, K-means Clustering Model, Random Forest Prediction Model.

1. Introduction

Since 1990, China's real estate market has been developing rapidly and housing prices have continued to rise, but frequent adjustments in property market policies have led to large fluctuations in housing prices. With the continuous acceleration of urbanization, the real estate market has become an important economic field in modern society. In this field, house price is a key indicator, which is of great significance to all parties such as home buyers, developers and government. Therefore, home price forecasting is very helpful for home buyers to choose the timing of home purchase and avoid the risk of asset depreciation. As home buyers are often at a disadvantage due to the asymmetry of home price information, home price forecasting can provide more reference information to help them assess the actual value of their homes and choose a more cost-effective property.

This paper quantifies the data of each characteristic of house price information and finds the influence of each factor on house price, then classifies the quantified house data, and explains the reasons and results of the classification. Finally, the index data from the above analysis are combined to build a follow-in forest house price prediction model.

2. Data Pre-processing

In this paper, we obtained a large number of feature values and data based on the house price information of Guangzhou city, including 'houseTotalMoney', 'houseSinglePrice', ' houseLocation', 'houseType','houseFloor', 'houseBuildingArea', 'houseStructure', 'houseInnerArea', 'houseOrientation', 'houseDecoration', 'houseElevatorRatio', 'houseElevator', 'reset cycle', 'houseTradeProperty', 'houseUsage', 'houseAgeLimit', 'housePrivilegeProperty', 'housePledge', etc. In subsequent calculations houseTotalMoney is represented by y and the rest of the data categories are represented by $x_1 - x_{17}$.

Upon inspection, we found that the data we obtained had more missing values and outliers. The processing of them is shown in Table 1.

Table 1. Missing and outlier handling

Processing Fields	Original data style	Processing	Number of bars processed
houseTotalMoney houseSinglePrice	The former is no data available, the latter is empty	Delete	195
houseInnerArea	East, south and other directions	Delete	5
houseStructure	No data available	Delete	25
houseInnerArea	No data available	Replace with data from houseBuildingArea	836

We then further quantify some of the feature values.

HouseLocation: Establish a mapping between categories

and values. Downtown = 3, peri-urban = 2, and distant = 1. To define the distance, the economic center of the district where each house is located is taken as the center of the circle, and then the distance of each house neighborhood location from the nearest economic center is calculated. Trisections and sextiles are used in statistics to describe the distribution of data, so when distinguishing between downtown, peri-urban and distant suburbs, the data set is divided by calculating the trisections and sextiles of the data set.

HouseStructure: The rating is based on the value and condition of the structure type. E.g. Staggered = 4, Duplex = 3, Flat = 2, Leap = 1.

HouseOrientation: Based on lighting and sunlight exposure, the eight directions of house orientation in Guangzhou were rated by reviewing the data. Southeast=8, South=7, Southwest=6, East=5, West=4, Northwest=3, North=2, Northeast=1. For houses with multiple orientations, the scores were averaged.

HouseElevatorRatio: Elevator ratio = number of elevators / number of households.

HouseElevator: With elevator = 1; without elevator or no data = 0.

HouseDecoration: Rating according to the degree of decoration: "other" is set to 0 because it cannot be judged. Other=0, rough=1, simple=2, fine=3.

HouseAgeLimit: The variable has two kinds of values: full five years and full two years, so construct dummy variables: the substance of the constructed dummy variables is full five years = 1, not full five years = 0, but in this question the value of not full five years is only full two years, so it is constructed as full five years = 1, full two years = 0.

HouseUsage: The variable has two values, commercial and residential, which indicate the type of use of the house. Therefore, the dummy variable is constructed as: commercial/residential = 0, ordinary house = 1.

HouseTradeProperty: The rating is based on the market price of the nature of the transaction: by consulting the data [1], the rating can be set as follows: commercial housing = 10, housing reform and private property = 8, affordable housing = 6, relocated housing = 4.

HousePrivilegeProperty: The variable has only two values, so dummy variables are constructed: non-common = 0, common = 1.

HouseType: Sums the number of rooms, kitchens and bathrooms of each house.

HousePledge: According to the security of the mortgage, there are four types of mortgages: CPF mortgage is the safest, commercial bank mortgage is the second, other mortgages are higher risk, and no mortgage is the highest risk, so the score can be set as follows: CPF mortgage = 3, commercial bank mortgage (agricultural bank, finance company, etc.) = 2, other mortgages (owner repayment, etc. = 1), no mortgage = 0. Add up.

HouseFloor: The median 11 of the middle floor is taken as the benchmark, avoiding the subjectivity brought by choosing any specific floor, and being more objective. The benchmark score is set according to the floor height type: high floor = 10, medium floor = 8, low floor = 6. Within each type, additional points are set according to the difference between the specific floor and the base floor. Taking the middle floor as the benchmark, each benchmark floor above the middle floor will add 0.5 points, and each floor below the middle floor will subtract 0.5 points.

Define variable reset cycle: Reset cycle = HouseListDate - HouseLastTrade. reset cycle is the time difference between the current listing time of the house and the time of the last trade. This time difference can reflect the length of time the house has been on the market, and can also indirectly reflect the appreciation or depreciation in the value of the house.

3. Stepwise Regression Model

Since the magnitudes of the individual factors differed significantly, the paper first standardized the quantified data. The correlation between the factors was initially explored and the Person coefficient matrix was calculated, and the results are shown in Figure 1.

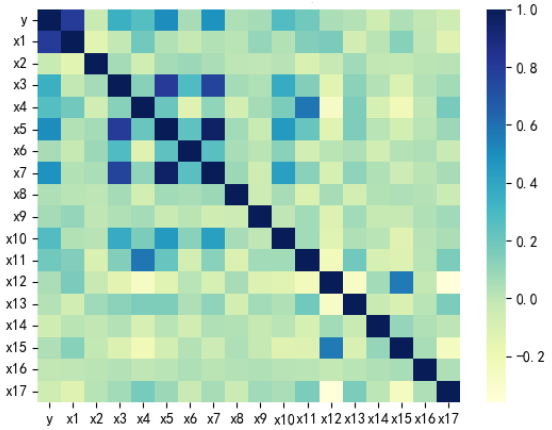

Figure 1. Heat map of correlation coefficient matrix

From the figure: $x_1, x_3, x_4, x_5, x_6, x_{10}, x_{11}$ on y are above 0.4. To further determine more accurately the influence of each factor on house prices, a stepwise regression model was established [2].

The process of selecting variables by stepwise regression method consists of three steps: firstly, weeding out the tested insignificant variables from the regression model, secondly, introducing new variables into the regression model, and thirdly, repeating step two over and over again until no new variables can be introduced.

To explore the effect on the dependent variable house price, for the 17 independent variables of the target, given a significance level of α_1=0.1 for the introduced variables and α_2=0.15 for the excluded variables, step two was repeated continuously and ten independent variables were introduced to obtain the stepwise regression model.

$$\begin{aligned} y = {} & 0.7937x_1 + 0.0459x_2 - 0.0649x_3 \\ & + 0.0288x_4 + 0.528x_5 \\ & - 0.0477x_6 + 0.0357x_{10} \\ & - 0.0522x_{11} - 0.0694x_{14} \\ & - 0.0193x_{15} \end{aligned} \tag{1}$$

Where the adjusted R^2 = 0.883 and the p-values of the t-test for each variable are shown in Table 2.

Table 2. T-test P-value table

Variables	P-value
x_1	0
x_2	0
x_3	0
x_4	0.0026
x_5	0
x_6	0
x_{10}	0
x_{11}	0
x_{14}	0
x_{15}	0.013

From Table 2: P<0.05 for each of the above variables, so the effect of the variables is considered significant.In summary: the effects of $x_1, x_2, x_3, x_4, x_5, x_6, x_{10}, x_{11}, x_{14}, x_{15}$ on house prices are considered significant and correlated, while the remaining variables have insignificant effects on house prices and are not strongly correlated.

4. K-means Clustering Model

4.1. Data Selection

Above, by stepwise regression, it was determined that with 'houseTotalMoney', 'houseSinglePrice', 'houseBuildingArea', 'houseUsage', 'houseStructure', 'houseLocation', 'houseElevatorRatio', 'houseType', 'houseElevator', 'houseFloor', 'houseAgeLimit' and other 11 reserved variables are used as data sets to classify this high-dimensional data using K-means clustering model.

4.2. Model Principle

Firstly, the elbow rule is used to determine the optimal number of clusters k for K-means clustering [3]. The core idea of the elbow rule is that as the number of clusters k increases, the sample division will be finer, the degree of aggregation of each cluster will gradually increase, and then the error squared and SSE will naturally become smaller gradually. And, when k is less than the true number of clusters, the decrease of SSE will be large because the increase of k will significantly increase the degree of aggregation of each cluster, and when k reaches the true number of clusters, the return of the degree of aggregation obtained by increasing k again will rapidly become smaller, so the decrease of SSE will plummet and then level off as the value of k continues to increase, which means that the graph of the relationship between SSE and k is an elbow of shape, and the value of k corresponding to this elbow is the true number of clusters of the data.

4.3. Specific Modeling

(1) K-means clustering is performed for a certain range of k values.

(2) Calculate the clustering evaluation index SSE (sum of squared errors) corresponding to each k-value, where：

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \quad (2)$$

C_i is the i-th cluster, p is the sample points in C_i, m_i is the center of mass of C_i (the mean of all samples in C_i), and SSE is the clustering error of all samples, which represents the clustering effectiveness.

(3) Draw the curve of k-value versus evaluation index.

(4) Find an obvious 'elbow' in the curve, where the k-value is usually the better clustering number.

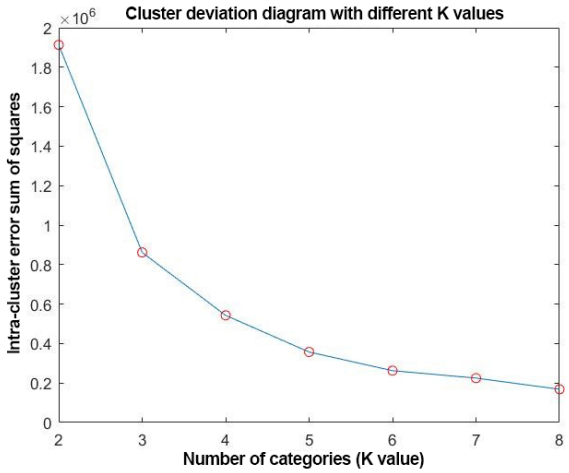

Figure 2. Cluster deviation diagram with different K values

As in Figure 2, when k=2, the SSE is large and a cluster contains a large number of dissimilar samples; as k increases, the SSE decreases rapidly and the samples are gradually divided into more suitable clusters; when k=4, 5, 6, 7, the decline of SSE slows down, and adding a cluster at this time does not bring about a large decrease in SSE; therefore, the cluster deviation plots for different k values obtained by the elbow rule know that when k = 3, the degree of distortion is substantially improved, and k = 3 is selected as the number of clusters.

The housing data were classified by iterating and then using the K-means clustering algorithm. The classification results are shown in Figure 3.

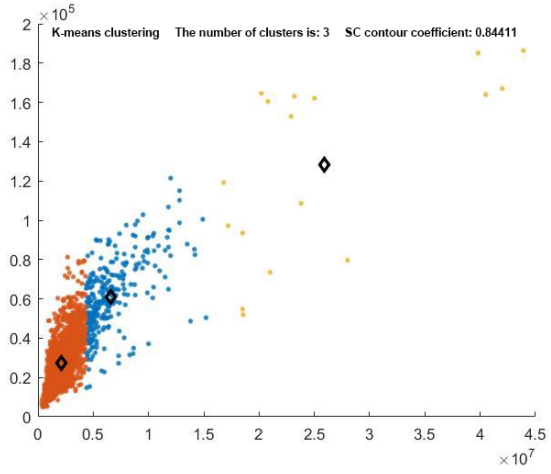

Figure 3. K-means clustering results graph

For the clustering points in red, this cluster contains the largest number of samples, which is about 65% of the total samples. This indicates that in the 11-dimensional feature space, this part of samples shows high similarity in the two key features of houseTotalMoney and houseSinglePrice; the sample points in this cluster are closely clustered together and clearly distinguishable from the blue and yellow clusters, which indicates that the clustering algorithm correctly captures the similar aggregation structure between samples and classifies This indicates that the clustering algorithm correctly captures the similar aggregation structure between samples and classifies the similar samples into the same cluster; it can be presumed that this cluster represents a clearer sample category, and this part of samples may belong to the same sample category in practical applications.

For the blue cluster: this cluster contains the middle number of samples, accounting for about 30% of the total samples. This indicates that the distribution of these samples in the 11-dimensional feature space is more dispersed, and the clustering algorithm cannot effectively capture the aggregation structure. This suggests that the key variables houseTotalMoney and houseSinglePrice play a role in distinguishing the samples in this cluster, but not as much as in distinguishing the samples in the red cluster.

For the yellow cluster: this cluster contains the least number of samples, about 5% of the total samples. As can be seen from the figure, the distribution of sample points in this cluster is more dispersed, and the separation from the red and blue clusters is not significant. This indicates that the K-means clustering algorithm is influenced by the discrete values and fails to find and use the aggregation structure well, dividing the samples that should be classified into the same cluster into small clusters of moderate size, which makes the clustering results less effective in prediction. However, the samples in this cluster still show some similarity in two key variables, houseTotalMoney and houseSinglePrice, resulting in this small cluster.

Also, as can be seen from the figure, the contour coefficient is 0.8411, which indicates that the clustering results obtained by the K-means clustering algorithm are relatively good, and the samples have a high similarity within the same cluster, while the differences between different clusters are relatively large. This may indicate that the clustering algorithm captures the similarities and differences between the samples better when dividing them into different clusters and obtains a more reasonable clustering structure.

Based on the two key variables, houseTotalMoney and houseSinglePrice, and the characteristics such as the range of quantities after clustering, the classification results are given as shown in Table 3.

Table 3. Classification results

House type	Category explanation	Clustering points
Economical housing	The total and unit prices of this category are in the lower range, making it an affordable option for the average family.	Red clustering points
Comfortable Housing	The total price and unit price of this type of housing are medium, neither low nor high end, and the living environment and facilities should be relatively comfortable and moderate.	Blue clustering points
High-end type housing	The total price and unit price of this category are in the higher range, and the living environment, decoration and facilities should reach a higher level of luxury.	Yellow clustering points

5. Random Forest Prediction Model

5.1. Model Principle

The basic idea of the random forest regression prediction algorithm is to achieve prediction by constructing multiple decision trees and using randomization to enhance the generalization ability and robustness of the model [4].

The training process of random forest regression prediction first extracts the training dataset from the original dataset by self-sampling method (bootstrap) and randomly selects a portion of features as candidate division features, and then constructs a decision tree for the training dataset using CART decision tree algorithm. For each tree, the splitting features of the nodes are selected from the set of candidate features and split by minimizing the objective function such as the mean square error on the dataset. Ultimately, the random forest prediction model is integrated from multiple decision trees. When making predictions, test data are fed into the random forest prediction model, each decision tree gives a prediction result, and the final prediction result is the average or weighted average of the prediction results of all decision trees.

5.2. Model Building

The quantified data were divided into training and test sets, 80% of the data were taken as the training set and 20% as the test set, the training and test sets were randomly selected each time, the data from the training set were normalized, and the model worked best when 30 decision trees were obtained by adjusting the parameters, and the random forest regression prediction model was constructed using the built-in function TreeBagger in MATLAB to predict the test set, and the predicted values and importance estimation plots were derived [5].

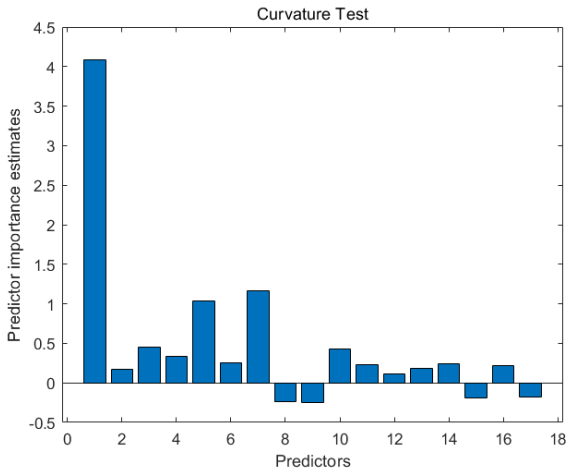

Figure 4. Importance estimation graph of raw data

Analysis of Figure 4 shows that the importance of each variable deviates significantly, and the model needs to be

processed by eliminating the variables with the least importance in turn until the importance of each variable is acceptable. After performing the variable elimination, the final model and the importance estimation plot were obtained (as shown in Figure 5).

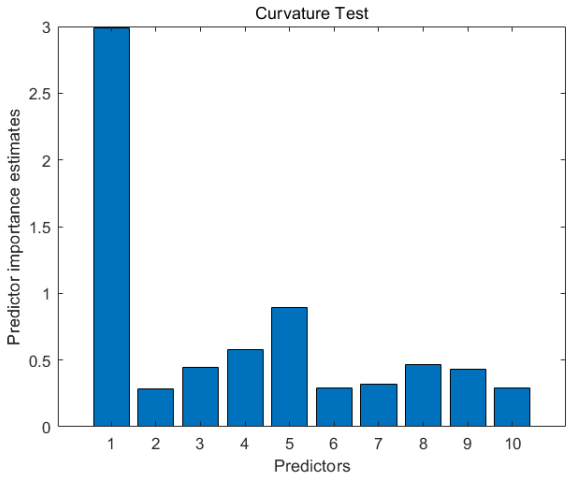

Figure 5. Importance estimation graph after removing variables

After removing the variables, the indicators $x_1, x_2, x_3, x_4, x_5, x_6, x_{10}, x_{11}, x_{14}$, and x_{15} were selected to construct the random forest prediction model, where R^2=0.92 and the model fit accuracy was high.

6. Conclusion

6.1. The establishment of simulation model

According to the analysis results of this paper, house prices are mainly affected by houseSinglePrice, houseLocation, houseType, houseFloor, houseBuildingArea, houseStructure, houseDecoration houseElevatorRatio, houseTradeProperty, houseUsage, which are 10 indicators that the government can refer to in order to formulate relevant housing policies to promote the stable and healthy development of the real estate market. Home buyers and investors can refer to these indicators to choose the right type of housing and investment direction.

The government can also use the random forest forecasting model to predict the house prices and formulate corresponding policy measures. Housebuyers and investors can use Random Forest Forecasting Model to predict house prices and make more informed decisions. Households can use random forest prediction models to assess the value of their own homes and make decisions accordingly to get a better return on their investment.

References

[1] Minjie Bian.A study of the factors influencing the variability of housing prices among cities in China[D].Nanjing University of Science and Technology, 2021. DOI : 10.27241/ d.cnki.gnjgu. 2021. 002031.

[2] Chunli Peng.Research on the factors influencing the consumption level of urban residents in Guangdong Province[J].China Collective Economy, 2023(09):24-27.

[3] Simei Lin,Huaguo Huang,Ling Chen.Combining random forest and K-means clustering to evaluate the severity of wetland fires[J].Remote Sensing Information, 2019,34(02):48-54.

[4] Yanlin Xu. Random Forest Model-based Bulk Evaluation of Used Residential Home Prices[D].Zhongnan University of Economics and Law, 2021. DOI: 10.27660/ d.cnki.gzczu.2021.002222.

[5] Fei Ling,Yanan Li.Integrated learning algorithm based house price prediction model[J].Information & Computer, 2022, 34(22): 96-100.