An empirical analysis of second-hand house transaction prices based on machine learning

Yi Cai*

Department of Finance, Capital University of Economy and Business, Beijing, China, 100071

*Corresponding author: contactAndye0605@163.com

Abstract. The Beijing real estate market is one of the most developed and representative real estate markets in China. As of May 25, 2016, Beijing accounted for 88.2% of the market residential transactions, and the Beijing property market has fully entered the era of second-hand houses. Therefore, the price of second-hand houses in Beijing can be judged by studying the factors influencing the price of second-hand houses in Beijing, and the study has theoretical significance. Taking Beijing second-hand houses as the research object, the data source is mainly designed as a crawler to mine the data and obtain the transaction attribute information of chain house network, and the data sample size is 2000 items in total [1]. In this paper, we obtained the important variables affecting the house price through multiple model analysis and predicted the data of 100 unknown house prices.

Keywords: Decision tree model; SVM model; Random Forest model; Adaboost model.

1. Introduction

1.1 Data source

The research object is Beijing second-hand houses, and the data source is mainly designed crawlers to capture the transaction listings on Chain Home. Data mining technology can effectively solve this problem by being able to use statistical data software and corresponding program algorithms that can categorize incomplete and arbitrary data information in a scientific and orderly manner according to a systematic procedure [2]. The sample size of the data is 2000 items. The reasons for choosing Chain Home are: First, Chain Home provides transaction The first reason is that Chain Home provides transaction data query, rich data type, and contains more important factors affecting the house price, which meets the research and analysis needs. Secondly, according to the "Survey Report on Consumer Satisfaction in China's Real Estate Service Industry", Chain Home has a good reputation for its authenticity of listing information and considerate service. Second, according to the "China Real Estate Service Industry Consumer Satisfaction Survey Report", Chain Home ranked first in satisfaction with the authenticity of listing information and considerate service.

1.2 Variable Description

Tu Jin et al. (2021) argued that policy factors, location factors, and housing attribute factors affect house prices [3]. Yang Hui et al. (2019) argued that factors such as supply factors, demand factors, land factors, and macro environmental factors affect house prices [4]. Cui Chengying (2011) argued that factors influencing the price of commercial housing in Beijing include consumer factors, real estate agent factors, and government factors [5]. Hong, Goho et al. argued that factors affecting house prices should start from factors affecting real estate supply and demand [6]. Combining the above literature and data availability, the following variables were selected. Quantitative variables include transaction period, number of price adjustments, number of people following, number of views, number of bedrooms, number of living rooms, number of kitchens, number of bathrooms, total number of floors, floor area, set area, year of completion, stairs, households, ratio of stairs to households, Baidu longitude, Baidu latitude, resident foreign population, resident population, total population, resident population density, a total of 21 quantitative variables. Qualitative variables include 14 qualitative variables, such as floor location, household structure, building type, house orientation, decoration, building structure, heating method, ownership age, equipped with elevator,

transaction ownership, house use, house age, house right ownership, and district. The above variables are explanatory variables and the explanatory variable is house price per square meter.

1.3 Data Description

Descriptive statistics for quantitative variables using the describe function in R. Specific metrics of interest include, mean, standard deviation, median, minimum, maximum, skewness, and kurtosis. Large standard deviations for transaction period, number of followers, number of views, and floor area may provide important information for modeling, but may also result from not standardizing the variables.

For qualitative variables taken, descriptive statistics in the form of bar charts are adopted. The descriptive statistics show that the second-hand houses basically present medium-high floors, flat floors, slab buildings, centralized heating, 70 years of property rights, steel-composite structure of buildings, commercial houses, ordinary residential houses for housing purposes, and full five years of property rights.

2. Data pre-processing

2.1 Missing value judgment and processing

In this paper, missing values are handled as follows: in case of missing, for one of the quantitative variables, delete the corresponding row; for another quantitative variable, replace it with the median; for another qualitative variable, replace it with the plural. Missing qualitative variables: floor, house type structure, building type, house use, house ownership, district, Baidu latitude and longitude (replaced by plural) Missing quantitative variables: price per square meter (deleted), transaction period, number of followers, number of views, building area (replaced by median). After data processing, there are no missing values.

2.2 Outlier judgment and handling

The outliers are judged and handled by observing the outliers with box line diagrams, mainly based on each district and county, and drawing a box line diagram of price per square meter for each district and county. And find the serial number where the outliers are located. Since the main purpose of this paper is to screen important variables, and supervised learning, the information contained in the outliers is still valuable, and no special treatment is done.

2.3 Standardized conversion of quantitative variables

To prevent data analysis from being caused by inconsistencies in the units of the data, all quantitative variables were standardized.

2.4 [0,1] transformation of qualitative variables

The specific rules for converting qualitative variables to 0 or 1 are that for dichotomous variables, only a set of 0 and 1 variables need to be converted.

2.5 Other

In the follow-up study, there are some need methods that require the target variable to be a dichotomous variable. Therefore, it is possible to transform house prices into [0,1] variables. Those below the house price mean are set to 0 and those above the house price mean are set to 1.

3. Preliminary analysis of the importance of variables

Finding predictor variables with better predictive power for the target variable and eliminating redundant variables to reduce model complexity, improve model explanatory power and predictive

warp. For the preliminary analysis of variable importance, a total of three common ideas are included, which are: without considering the target variable and the algorithm, based only on the characteristics of the independent variable itself; without considering the algorithm, based on the relationship between the independent variable and the target variable; related to the target variable and the algorithm, the judgment of the importance of the independent variable takes into account the target variable and the prediction algorithm used. In this paper, four methods are used for preliminary analysis of variable importance: ANOVA, chi-square test, AUC method, and Relief method. ANOVA is applied to quantitative independent variables and dependent variables; chi-square test is applied to qualitative independent variables and dependent variables; AUC method is applied to quantitative independent variables and qualitative dependent variables; and Relief method is a general rule with no significant data requirements. In the variable importance analysis, the results of the above four methods on screening variables will be combined to determine the important factors affecting the price per square meter.

For the qualitative variables, the chi-square test and Relief's rule show that the qualitative variables that have significant effects on the dependent variable include: floor, house type structure, building type, house orientation, decoration, building structure, heating method, ownership age, equipped with elevator, cross-ownership, house use, house age, house ownership and district. For the quantitative variables, the variables that have a significant effect on the dependent variables, as shown by ANOVA and AUC rule, include: number of people of concern, floor area and proportion of staircase households

Among the above qualitative and quantitative variables, there are factors that have an indirect effect on the house price, such as the number of people who pay attention to the house, and there are factors that have a direct effect on the house price, such as the floor, the structure of the house, the building type, and the orientation of the house. The above conclusion is in line with the knowledge of economics and common sense of life, and will focus on the above variables in the following analysis. However, in order to obtain accurate information on house prices, only individual variables (such as Baidu longitude, Baidu latitude, population density, etc.) are deleted, and not all variables that have a weak influence on house prices are deleted.

4. Model construction and solving

Since the data requirements of the SVM model are that both the independent and dependent variables are categorical variables, and the decision tree model does not have obvious data requirements. Also, in order to facilitate comparison between models a uniform test and validation set needs to be selected, so the datasets selected for the following models are all datasets in which the qualitative variables are converted to categorical variables (the same dataset). This dataset is consistent with the previous results after data preprocessing, with only variable censoring manipulated.

Split the dataset in a 3:1 ratio and save it. Ensure that the datasets used in subsequent models are the same dataset to facilitate comparison of model correctness.

4.1 Decision tree model based on Gini index

In machine learning, a decision tree is a predictive model; it represents a mapping relationship between object attributes and object values. Each node in the tree represents an object, each bifurcated path represents a possible attribute value, and each leaf node corresponds to the value of the object represented by the path from the root node to that leaf node. A decision tree has only a single output; if you want to have a plural output, you can build a separate decision tree to handle different outputs. Decision trees are a frequently used technique in data mining for analyzing data and also for making predictions [7].

The decision tree model consists of the following steps: building the tree, outputting the relevant rules, outputting the results of the training and test sets, splitting the wrong sample numbers, and the

ROC curve. The decision rules are: first, determine whether the number of the district is greater than or equal to 9, if it is greater than 9, then see whether it was built before 2000, if it is before 2000, then it is a low-price house. If the district number is less than 8, then judge whether the district number is greater than or equal to 6, if it is greater than or equal to 6, then it is a high price house. If the district number is less than 6 and also less than 4, it is a high price house. If the district number is less than 6 but greater than or equal to 4, determine whether it is centrally heated, and if it is centrally heated, it is a low-priced house. If it is not centrally heated, then see if the bedroom number is greater than greater than 2, if it is less than 2, it is a high house price; if it is greater than 2, then see if the ladder number is less than 2, if it is less than 2, it is a low house price and vice versa.

The important influencing variables are the most important variables including the district where it is located, the year it was built, the proportion of stairs, the floor area, the household, and (whether it is) the slab building.

The correctness of the test set is as follows; the probability of predicting class 0 with a sample of class 0 is 91.04%; the probability of predicting class 1 with a sample of class 1 is 94.79%. The overall correct rate is 0.92, which is a high overall correct rate. the value of AUC indicates how good the classifier is, and a larger AUC value represents good performance. The larger AUC under this method is 0.970, so the decision tree is well constructed.

4.2 Decision tree model based on entropy value

The decision rule is: first judge whether the number of the district is greater than or equal to 9, if it is greater than 9, then see whether the completion date is before 2000, if it is before 2000, then it is a low-price house. If the district number is less than 8, then judge whether the district number is greater than or equal to 6, if it is greater than or equal to 6, then it is a high price house. If the district number is less than 6 and also less than 4, it is a high price house. If the district number is less than 6 but greater than or equal to 4, determine whether it is centrally heated, and if it is centrally heated, then it is a low-priced house. If it is not centrally heated, then see if the number of bedrooms is greater than or equal to 2. If it is less than 2, it is a high house price; if it is greater than 2, then see if the ratio of staircases is greater than or equal to 1.8. If it is less than 2, it is a high house price, and vice versa is a low house price.

The important influencing variables include: the district where it is located, the year it was built, the proportion of terrace households, households, floor area, slab floor, and the number of bedrooms.

The correctness of the test set is as follows; the probability that the sample is class 0 and the prediction is class 0 is 91.75%; the probability that the sample is class 1 and the prediction is class 1 is 90.63%. The overall correct rate is 0.92, which is a high overall correct rate. the value of AUC marks the goodness of the classifier, and a larger AUC value represents good performance. The larger AUC under this method is, 0.971, so the decision tree is well constructed.

4.3 SVM model

SVM is a two-class classifier. Its basic model is defined as a linear classifier with maximum interval on the feature space, and its learning strategy is interval maximization, which can eventually be translated into the solution of a convex quadratic programming problem [8]. The parameters of SVM are kernel function is linear, cost=10, gamma=1.

The correctness of the test set is as follows; the probability that the sample is class 0 and the prediction is class 0 is 95.34%; the probability that the sample is class 1 and the prediction is class 1 is 42.71%. The overall correct rate is 0.82, which is a high overall correct rate. the value of AUC marks the goodness of the classifier, and a larger AUC value represents good performance. The larger AUC under this method is 0.690, so the decision tree is well constructed.

4.4 Random Forest Model

In constructing a random forest, which consists of multiple decision trees, each of which is not the same, we put back a random selection of samples from the training data and do not use all the features

of the data, but rather randomly select some of the features for training [9]. Each tree uses different samples and features, and the results at training are different. The reason for this is that before starting the training, it is not possible to know which part of the data has anomalous samples and which features best determine the classification results, and the random process reduces the impact of both influences on the classification results.

The relevant parameters of the random forest model are: ntree=500, mtry=3, etc., and set the output important influence variables. The significant variables include the number of bathrooms, (whether it is) leapfrog, (whether it is) tower, (whether it is-) slab, house orientation, (whether it is) steel-composite structure, (whether it is) brick-composite structure, number of ladders, (whether it is) Class II affordable housing, (whether it is) single-apartment housing, (whether it is) ordinary housing.

The correctness of the test set is as follows; the probability that the sample is class 0 and predicted to be class 0 is 98.92%; the probability that the sample is class 1 and predicted to be class 1 is 38.54%. The overall correct rate is 0.83, and the overall correct rate is high. the value of AUC marks the goodness of the classifier, and a larger AUC value represents good performance. The larger AUC under this method is 0.687, so the decision tree is well constructed.

4.5 Adaboost model

AdaBoost (Adaptive Boosting) is an effective and practical Boosting algorithm that sequentially trains a weak learner in a highly adaptive manner [10]. For the classification problem, the AdaBoost algorithm adjusts the weights of the data based on the previous classification, with the weights of the misclassified samples in the previous weak learner being increased and the weights of the correctly classified samples being decreased in the next weak learner, and a new weak learner is added to the model in each iteration. The weights are repeatedly adjusted and the weak learners are trained until the number of misclassifications falls below a predetermined value or the number of iterations reaches a specified maximum, resulting in a strong learner. In short, the core idea of the AdaBoost algorithm is to adjust the weights of the misclassified samples and iteratively upgrade them.

Adaboost parameters are set as follows: mfinal=60, maxdepth=4, maxcompete=6, cost=error_cost. etc. The important variables include district, number of views, year of completion, total number of floors, unit area, days of transaction cycle, number of followers, floor area, and proportion of households and staircases.

The correctness of the test set is as follows; the probability of predicting category 0 for a sample of category 0 is 96.06%, and the probability of predicting category 1 for a sample of category 1 is 90.63%. The overall correct rate is 0.95, which is a high overall correct rate, and the value of AUC indicates the goodness of the classifier. The large AUC value of 0.690 in this method means that the decision tree is well constructed.

4.6 Model Comparison and Description

By comparing the class 0 correct rate, the correct rate from high to low is random forest model, Adaboost model, SVM, decision tree 2, and decision tree 1. The class 1 correct rate from high to low is: decision tree 1, decision tree 2=Adaboost, SVM, random forest model. The total correct rate from high to low is: Adaboost, Decision tree 1=Decision tree 2, Random Forest model, SVM. the AUC from high to low is Decision tree 2, Decision tree 1, SVM=Adbboost, Random Forest model.

In the decision-making process, the model with higher total correct rate is better, so the Adaboost model is the best among the above models. This model is also used in the subsequent predictions. Comparison of model correctness and AUC and importance of variables is shown in Table 1

Table 1. Comparison of model correctness and AUC and importance of variables

Models	0 class correct rate	1 class correct rate	Total correct rate	AUC	Variables used and their importance (from largest to smallest)
Decision Tree 1	91.04%	94.79%	92%	0.970	District, year of construction, proportion of staircase, building area, household, (whether) slab building
Decision Tree 2	91.75%	90.63%	92%	0.971	District, year of construction, proportion of staircases, households, floor area, (whether) slab, number of bedrooms
SVM	95.34%	42.71%	82.00%	0.690	----------
Random Forest Model	98.92%	38.54%	83.00%	0.687	Number of bathrooms, (whether it is) leap floor, (whether it is) tower, (whether it is) slab floor, orientation of housing, (whether it is) steel mixed structure, (whether it is) brick mixed structure, stairs, (whether it is) Class II affordable housing, (whether it is) single apartment housing, (whether it is) ordinary housing
Adaboost model	96.06%	90.63%	95.00%	0.690	District, number of views, year of construction, total number of floors, unit area, days of transaction cycle, number of followers, floor area, households, ratio of staircases

5. Conclusion

By comparing the four models, it can be found that the important variables are the district where the house is located, the year of completion, the proportion of terrace households, the floor area, the households, the building type, the building structure and the ownership of the transaction.

In order, among the above factors, those that directly affect the house price are age of construction, proportion of terraced households, floor area, households, building type, building structure, and ownership of the transaction. The price of an older house may be lower because the quality is affected by time. The terrace ratio reflects the information of both the number of households and the number of stairs, and therefore affects the price. Building area is a direct influence on price. An increase in area will show an upward and then downward trend in price. Those with fewer households are likely to be villas and therefore have higher prices. The price will be higher if the type of construction is fine decoration than if the type of construction is simple decoration. The price will be higher if the building structure is steel-composite structure. The different ownership of the transaction will also affect the price.

The location of the district indirectly affects the price, because different districts mean different resources. Therefore, prices are higher in Haidian District and Xicheng District.

The selection of the above significant models is basically consistent with the conclusions reached using ANOVA, chi-square test, AUC method, and Relief's rule in the above section.

Adaboost was used for the extrapolation prediction because the Adaboost model has the highest total correct rate. 100 data were selected for prediction and the prediction results were as Figure 1

```
 [1] "1" "0" "0" "0" "0" "0" "0" "0" "0" "0" "0" "0" "0" "0" "1" "0" "0" "0"
[19] "0" "0" "0" "1" "1" "1" "1" "1" "0" "1" "0" "0" "0" "0" "0" "0" "0" "0"
[37] "0" "1" "0" "0" "0" "0" "0" "1" "0" "1" "0" "0" "0" "0" "0" "1" "1" "0"
[55] "0" "0" "0" "1" "1" "0" "0" "0" "0" "0" "0" "1" "0" "1" "1" "0" "1" "0"
[73] "0" "1" "0" "0" "0" "0" "0" "0" "0" "1" "0" "1" "1" "0" "0" "1" "1" "0"
[91] "1" "1" "0" "0" "0" "0" "1" "1" "0" "1"
```

Figure 1. Adaboost extrapolation prediction results

References

[1] Liu M et al. Application of data mining technology in the era of big data. Science and Technology Herald.2018, 36 (9): 73.

[2] Xue Yang. Analysis of data mining technology application in economic statistics. Quality and Market. 2022 (03): 184.

[3] Tu J et al. Differential research on the influencing factors of house prices in China's cities--analysis based on big data of second-hand house market in Chengdu. Price Theory and Practice. 2021 (10): 75 - 76.

[4] Yang, Hui et al. A study on the influencing factors and contribution of house prices in Chinese cities--an analysis of relative importance based on R^2. Exploration of Economic Issues. 2019 (11): 51 - 53.

[5] Cui Chengying et al. An empirical analysis of factors influencing commodity house prices in Beijing. Productivity Research. 2011 (09): 78 - 79.

[6] Hong, Gehao et al. Theoretical study of house price influencing factors. China Economic and Trade Journal. 2010 (02): 72.

[7] T. Y. Zhang et al. Decision tree algorithm for big data analysis.2016 (6A): 375.

[8] Liu F.Y. et al. A review of support vector machine models and applications.2018, 27 (04): 1.

[9] Li Huanxu et al. Random forest-based prediction and analysis of second-hand house prices in Shenzhen. Modern Information Technology. 2021 (15): 101.

[10] Cao Y et al. Research progress and prospects of AdaBoost algorithm. Journal of Automation.2013, 39 (06): 745.