Research on Housing Price Forecasting Model Based on Multiple Linear Regression Model and Neural Network Model

Ruihong Xu*
1720759381@qq.com

Zhengzhou University

Abstract——There is a huge demand for the housing market in China. Effective forecasting models can provide reference for potential house buyers, thus avoiding blind purchase. In addition, policymakers can adjust real estate policies based on model predictions to reduce the stagnancy of the real estate markets, prevent the occurrence of economic crisis and maximize the benefits1.[1] Based on the second-hand housing data of Bao 'an District, Shenzhen city, this paper establishes multiple linear regression model and neural network model respectively, and tests the model with test set data, and obtains the conclusion that the neural network model is better in the accuracy of describing experimental data and fitting effect. Besides, this paper studies the role of multiple linear regression models in helping consumers to buy houses on demand and in helping real estate practitioners to set house prices according to influencing factors. The multiple linear regression model can help consumers and real estate practitioners to intuitively understand the factors that significantly affect housing prices. Furthermore, consumers can combine their own needs based on these factors, select the houses with conditions and prices that meet expectations; real estate practitioners can combine reality, develop an appropriate price strategy.

Keywords: multiple linear regression model, neural network model, housing price; forecasting.

1 INTRODUCTION

Housing price is always the focus of people's attention. People buy and sell second-hand houses mostly through housing intermediary. However, some real estate agents may artificially inflate the price in order to earn higher commission, resulting in the bubble phenomenon that the housing price does not match the actual price. It is of great benefit to consumers and real estate practitioners to accurately predict the price of houses based on certain influencing factors. Based on the second-hand housing data of Bao 'an District, Shenzhen city, this paper studies the housing price by establishing linear regression model and neural network model, and compares the two models for housing price prediction[2]. Furthermore, this paper discusses and studies the effect of the model on consumers buying houses and real estate agents setting prices. Furthermore, this article provides a way to help consumers choose suitable houses according to their needs, and to help real estate practitioners develop a reasonable price strategy based on actual conditions by judging whether the factors of concern have significant effects on prices.

ICEMME 2022, November 18-20, Nanjing, People's Republic of China

DOI 10.4108/eai.18-11-2022.2327165

2 MODEL INVESTIGATION

2.1Research methods

Multiple linear regression algorithm thought

Let the linear regression model of random variable Y and general variablex1,x2,…,xp be

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon \tag{1}$$

In the formula, β_0 , β_1 ,…, β_p are P+1 unknown parameters. β_0 is called the regression constant. β_1 ,…, β_p are called regression coefficients. Y is called the explained variable (dependent variable), x1,x2,...,xp is called an explanatory variable (independent variable). When $p \geq 2$, the above formula is called multiple linear regression model. ε is a random error, Generally assumed that [3]:

$$\begin{cases} E(\varepsilon)=0 \\ \mathrm{var}(\varepsilon) = \sigma^2 \end{cases} \tag{2}$$

Multivariate weighted least square method

For the general multiple linear regression model $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i$, $i = 1, 2, ..., n$ (3), when the error term ε_i exists heteroscedasticity, the sum of squares of weighted deviations is $Q_w = \sum_{i=1}^{n} w_i (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - ... - \beta_p x_{ip})^2$ (4). In the formula, wi is the weight of the observation number i given. The weighted least squares estimation is to find the estimates $\hat{\beta}_{0w}, \hat{\beta}_{1w}, \hat{\beta}_{2w}, ..., \hat{\beta}_{pw}$ of parameters β_0 , β_1 ,…, β_p so that Q_w becomes minimal. Denote as:

$$W = \begin{bmatrix} w_1 & & & \vdots \\ & w_2 & & \\ & & \ddots & \\ \vdots & & & w_n \end{bmatrix} \tag{5}$$

It can be proved that the matrix expression of the weighted least squares estimation is

$$\hat{\beta}_w = (X^{'} W X)^{-1} X^{'} W y \tag{6}$$[3].

Artificial neural network regression algorithm thought

The principle of neural network is to weighted average the value of the upper node to the lower node, and finally to the output layer node, and then feedback to the previous layer according to the error, and then re-weighted average, and so on repeatedly training until the error is within the allowable range. The following formula can illustrate the weighting process of a general neural network.

$$yi = f^*(\Sigma_k w_{kj} z_{kj} + w_{0j}) = f^*\{\Sigma_k w_{kj}[f(\Sigma_i w_{ik} x_i + w_{0k})] + w_{0j}\} \tag{7}$$

In the formula, wik is the weight of the node number k of the independent variable xi in the hidden layer; wkj is the weight of the node number K in the hidden layer to the dependent variable number J; zkj is the value of the node number k in the hidden layer. f^* and f here are the activation functions and are usually defined as an S-shaped logistic function[3]:

$$f(x) = \frac{1}{1+e^{-x}} \tag{8}$$

2.2 Introduction of data

The data set comes from the information of second-hand houses for sale in Bao'an District, Shenzhen, captured by SouFun.com. The data set contains the information of the number of rooms, the number of halls, area, floor, school district room, subway, and price of the house. Among them, the floors are divided into "high", "middle", and "low" (relative height), which are represented by two dummy variables high. floor and middle. floor. For example, if the floor is high, high. floor takes 1, and middle. floor takes 0. The school variable is divided into "is a school district room" and "not a school district room", then the feature "is a school district room" is taken as 1, and the feature "is not a school district room" is taken as 0. The subway variable is divided into two characteristics: "close to the subway" and "not close to the subway". The characteristic "close to the subway" is taken as 1, and "not close to the subway" is taken as 0. The experiment adopts an 8-dimensional feature vector, and the price feature is the target that the author wants to predict. The data set contains a total of 1251 data samples. The author uses the first 70% of the data samples as training samples and the last 30% as test samples. Some data examples of the housing price data set are shown in Table 1. After inspection, there is no missing data or outliers.

TABLE 1. SOME DATA EXAMPLES OF THE DATA SET

roomnum	3	4	2	2	3
hall	2	2	2	2	1
area/m²	89.3	127	48.9	49.07	104.03
school	0	0	0	0	1
subway	0	0	0	0	0
high floor	0	1	1	1	1
middle floor	1	0	0	0	0
Per price/(ten thousand yuan/m²)	7.0773	6.9291	7.2597	8.0497	5.2869

2.3Multiple linear regression model construction

Stepwise regression

The author selected the optimal subset through stepwise regression: roomnum, hall, area, school, subway, and used these as the independent variables, and per price as the dependent variable to establish a multiple linear regression model, which is recorded as the lm1 model. The results are shown in Table 2.

TABLE 2. LM1 MODEL RESULTS

variable	coefficient	t	P> | t |	Model accuracy
Intercept	5.307579	14.606	$< 2\times10^{-16}$	R^2=0.2608 Adjust R^2= 0.2565 F statistic p value < 2.2×10^{-16}
roomnum	0.351383	2.602	0.00942	
hall	-0.942344	-4.771	2.15×10^{-6}	
area	0.007537	2.168	0.03043	
school	0.738221	5.120	3.77×10^{-7}	
subway	1.706239	11.832	$< 2\times10^{-16}$	

The coefficients of the respective variables are significantly non-zero at the 0.05 level. And the P value of the F statistic is less than 2.2×10^{-16}, so the lm1 model is significant at the 0.05 level.

The author did a white test on the model, and got p-value = 3.46×10^{-15}, so it is believed that the model has heteroscedasticity. Further, the author used the multivariate weighted least squares method to modify the model and obtained a new model, which is recorded as the lm2 model. The results of the lm2 model are shown in Table 3.

TABLE 3. LM2 MODEL RESULTS

variable	coefficient	t	P> | t |	Model accuracy
Intercept	5.323501	397.30	$< 2\times10^{-16}$	R^2=0.993 Adjust R^2= 0.9929 F statistic p value < 2.2×10^{-16}
roomnum	0.352005	91.30	$<2\times10^{-16}$	
hall	-0.939435	-226.29	$< 2\times10^{-16}$	
area	0.007271	30.17	$< 2\times10^{-16}$	
school	0.738509	94.72	$< 2\times10^{-16}$	
subway	1.701273	197.71	$< 2\times10^{-16}$	

The coefficients of the respective variables are significantly non-zero at the 0.05 level. And the P value of the F statistic is less than 2.2×10^{-16}, so the lm2 model is significant at the 0.05 level. The lm2 model R^2 and adjusted R^2 are significantly increased compared with the lm1 model. The adjusted R^2 of the lm2 model reaches 0.9929, which means that the model fits well.

Model checking

The author uses the variance expansion factor method to diagnose the multicollinearity of the model. The results are shown in Table 4.

TABLE 4. VARIANCE EXPANSION FACTOR

variable	roomnum	hall	area	school	subway
VIF_j	4.110924	2.166994	5.254942	1.000346	1.219510

The variance expansion factors of all independent variables of the lm2 model are not greater than 10, and the average ($\overline{VIF} = \frac{1}{5}\Sigma_{j=1}^{5} VIF_j = 2.75$) is slightly greater than 1. Therefore, it is believed that the model does not have serious multicollinearity. After testing, there is no autocorrelation phenomenon among the random error terms of the model.

Final model

After continuous testing and optimization, the author got the final multiple linear regression model lm2 model:

$$per\_price = 5.3235 + 0.3520 roomnum - 0.9394 hall + 0.0073 area + 0.7385 school + 1.7013 subway$$

2.4Artificial neural network regression model construction

Building a neural network model

The author determined that the number of hidden layer nodes is 6, the weight attenuation parameter is 0.1, and the neural network model is established with roomnum, hall, area, school, subway as the independent variables, and per price as the dependent variable. Among them, the author sets the learning rate to 0.1, the number of iterations to 100, and the training algorithm is the Levenberg-Marquardt algorithm. The structure diagram is as follows.

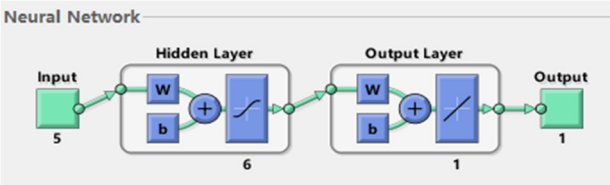

DIAGRAM 1 NEURAL NETWORK STRUCTURE DIAGRAM

2.5Model test results and model comparison

The author tested the two models with the test set data and the results are shown in Table 5.

TABLE 5. TEST RESULT

	R^2	MSE
Multiple linear regression model	0.9120	4.4446
Neural network model	0.9172	4.1661

From the test results, the goodness of fit of the multiple linear regression model is slightly lower than that of the neural network model, and the mean square error of the multiple linear regression model is higher than the mean square error of the neural network model, so the neural network model is better in the accuracy of describing experimental data and the effect of

fitting. In addition, the neural network model also has obvious advantages in the simulation and simulation capabilities of the initial data[5]. The linear regression model has advantages in studying whether the independent variable has a significant influence on the dependent variable.

2.6Model application

The housing price prediction model has played a certain role in helping consumers purchase houses on demand, and helping real estate practitioners to determine house prices based on influencing factors. According to the established multiple linear regression model, the number of rooms, the number of halls, the area of the house, whether it is a school district room, whether it is close to the subway and other factors have a significant impact on the housing price. The number of rooms and the number of halls are used to represent the shape and size of the residence. Its area has a direct impact on the degree of living comfort. In addition, the current urban infrastructure is not perfect, such as differences in school resources, and differences in the distribution of subways. These imperfections are the main reason for the difference in urban housing prices. Consumers can refer to the significant variables in the multiple linear regression model and combined with their own needs to measure the price growth brought about by various factors in the housing price, in order to choose a house that meets their expectations[6]. For example, consumers can choose a house in a suitable location according to their needs for the subway. Studies have shown that urban subways are an important factor influencing housing price changes along the lines, and that subways have a very obvious effect on real estate appreciation[7]. If consumers generally travel in the city by taking private cars instead of public transportation, then they can consider not buying houses along the subway line, thereby eliminating the effect of subways on house prices. Correspondingly, real estate practitioners can also consider factors that have a significant impact on housing prices, such as whether there is a subway nearby, whether it is close to a school, to decide whether to build a house, or to determine the price of a house based on these factors. For example, the subway has a significant impact on housing prices and the housing prices along the subway are related to time and distance. For real estate developers, they can develop the layout along the subway line before the subway construction opens, and do a good job of publicizing the real estate along the subway line. At the same time, according to the location and time of the real estate, they can analyze the degree of influence of the subway on housing price, and then predict the price trend, so as to formulate a reasonable pricing strategy[8].

3 CONCLUSION

In this paper, based on the second-hand housing data in Bao'an District, Shenzhen, a multiple linear regression model and a neural network model were established, and the two models were tested with the test set data, and the goodness of fit of the two models was 0.9120 and 0.9172 respectively. Further, the author conducted a comparative analysis on the prediction effects of the two models, and obtained the conclusion that the neural network model is better in the accuracy of describing the experimental data and the fitting effect, and explored how the multiple linear regression model can help consumers purchase houses on demand, and help real estate practitioners formulate house prices based on influencing factors.

In this paper, only a few factors are selected as variables among all the influencing factors of house prices, so it has certain limitations. In addition, this article only studies the single-prediction model. In terms of housing price prediction, the prediction accuracy of the single-prediction model is not ideal. In the future, the author can consider applying inherited learning to housing price prediction[9].

REFERENCES:

[1] Li, D., Liu, L., & Lv, H. (2021). Prediction of China's Housing Price Based on a Novel Grey Seasonal Model. Mathematical Problems in Engineering, 2021, 1–11. https://doi.org/10.1155/2021/5541233

[2] Ding Fei., & Jiang Ming. Yan. (2021). Housing price prediction based on improved lions algorithm and BP neural network model. Journal of Shandong University (Engineering Science Edition), 51, 1–9.

[3] He Xiao. Qun. (2017). Applied regression analysis (R language edition). Publishing House of Electronics Industry.

[4] Wu. (2015). Complicated Data Statistics Method-Application Based on R (The 3rd edition of the postgraduate books of institutions of higher learning) (1st ed.). China Renmin University Press.

[5] Zhang Jing. Yang., & Pan Guang. You. (2013). Comparison and application of multiple linear regression and bp neural network prediction models. Journal of Kunming University of Science and Technology (Natural Science Edition), 38, 1–7.

[6] Li, S., Jiang, Y., Ke, S., Nie, K., & Wu, C. (2021). Understanding the Effects of Influential Factors on Housing Prices by Combining Extreme Gradient Boosting and a Hedonic Price Model (XGBoost-HPM). Land, 10(5), 533. https://doi.org/10.3390/land10050533

[7] Yang Qing., & Wang Zi. (2015). Analysis of the influence of subway on housing prices along the line based on gray prediction. Henan Science, 33, 1–5.

[8] Yang Lu. Yao., Zhou Yan. Min., & Li Yi. Wen. (2019). The impact of Hefei subway on the housing prices along the line. Journal of Jiujiang University (Natural Science Edition).Published.

[9] Yang, B., & Cao, B. (2018). Research on Ensemble Learning-based Housing Price Prediction Model. Big Geospatial Data and Data Science, 1(1), 1–8. https://doi.org/10.23977/bgdds.2018.11001