ISSN: 2688-9323

Prediction of Second-hand Housing Prices in Chongqing based on Lasso Regression and BP Neural Network

Qi Gou

School of Sichuan Agricultural University, Yaan, China

gouqi2001@163.com

Abstract

With the continuous prosperity and development of the Second-hand housing market, accurate Second-hand housing price prediction can help people make correct decisions and protect people's legitimate rights as much as possible. In this paper, Lasso and BP neural network are used to predict the Second-hand houses price in Chongqing, and the prediction results are compared. In the two prediction models, the prediction error of BP neural network model is smaller, which shows that the prediction accuracy of BP neural network model for Second-hand house price is higher. It can be seen that BP neural network has certain advantages in dealing with housing price related data. In the future research, it can be applied to other fields for prediction research.

Keywords

Second-hand House Price Prediction; Lasso; BP Neural Network; Second-hand House.

1. Introduction

China's urbanization is accelerating with the rapid development of social economy, and the speed of real estate development and construction is faster and faster [1]. Both the impact of inflation and the increase of investment demand and speculative demand have promoted the growth of urban housing demand. The increasing demand for housing has led to the soaring price of new houses in recent years. Taking Chongqing as an example, the average transaction price of housing in 2022 has reached 11085 yuan every square meter, which not only makes the majority of urban low-income people unable to buy, but also makes it difficult for the working class and urban white-collar workers. Therefore, people have invested their purchase demand in "Second-hand housing" with high cost performance.The Second-hand housing market is booming [2]. However, due to the relatively complex transaction procedures of Second-hand housing, more relevant policies involved, and its market system needs to be developed and improved, there are still many problems that can not be ignored. According to statistics, nearly 80% of Second-hand houses are sold through third-party intermediaries, but many intermediaries trade in the way of "buying low and selling high", disturbing the market order and price. In view of this, there is an urgent need for a more efficient and fair way to disclose the housing information and objectively evaluate the house price. Therefore, this paper will select the transaction price of Second-hand housing market in Chongqing for prediction research.

There are many ways to predict house prices. Ren Shijie introduced hedonic characteristic price model and BP neural network into the improvement of market method, and constructed the real estate price evaluation model along the subway in Chengdu. Through experiments, it was found that the accuracy of the model reached 95.21%, and the prediction level was very significant [3]; Wang Jingxing used 79 factors affecting house prices to predict house prices. Using stacking model, integrating Lasso regression model and xgboost regression model, and using the 50% discount cross validation scheme, he obtained an RMSE of 0.128, indicating that

ISSN: 2688-9323

the model is good [4]; Weldensie T. Embaye predicted the rental value of houses in household surveys in Tanzania, Uganda and Malawi. The results show that machine learning method is the best model to predict house value using out of sample data sets in all countries and years [5].

Machine learning is a prediction method with high efficiency, good accuracy and strong adaptability [7]. In this paper, Lasso and BP neural network are used to predict the Second-hand house price in Chongqing from 2018 to 2020, and the error analysis is carried out for the prediction results of Second-hand house price, in order to obtain the optimal prediction model.

2. Method

2.1. Lasso Regression

Lasso regression model has great advantages in solving collinearity and over fitting problems, and the model can also select variable characteristics [6]. Lasso method is based on reducing the set of variables. By constructing a penalty function, it can compress the coefficients of variables and make some regression coefficients become 0, so as to achieve the purpose of variable selection. Therefore, the objective function of Lasso regression model can be expressed as:

$$\beta^{Lasso} = \arg_{\beta} \min[\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j|] \tag{1}$$

The formula is also equivalent to:

$$\begin{cases} \arg_{\beta} \min \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 \\ s.t.\sum_{j=1}^{p}|\beta_j| \le t \end{cases} \tag{2}$$

Among them, $\sum_{j=1}^{p}|\beta_j|$ is the penalty term of the function, λ is the penalty coefficient.An optimal value is estimated iteratively.

2.2. BP Neural Network

Neuron model is designed to simulate the structure of biological neurons. The typical neuron structure is shown in Figure 1 below:

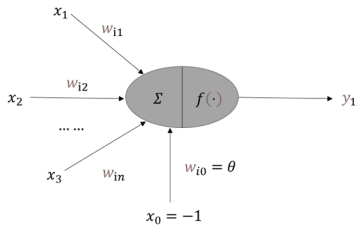

Figure 1. Typical neuronal structure

The full name of BP neural network is called back propagation neural network, which belongs to supervised learning algorithm and is often used to train multi-layer perceptron. Because BP neural network successfully solves the weight adjustment problem of multilayer feedforward neural network for solving nonlinear continuous function, nearly 90% of the neural network models in the practical application of artificial neural network adopt BP neural network and its variation form [8]. BP neural network has three or more layers of structure, namely: input layer, hidden layer and output layer. The hidden layer can have one or more layers. The neurons

ISSN: 2688-9323

between layers are fully connected, and there is no connection between neurons in layers. The nodes of each hidden layer generally use sigmoid function. The structure of BP neural network is shown in Figure 2 below:

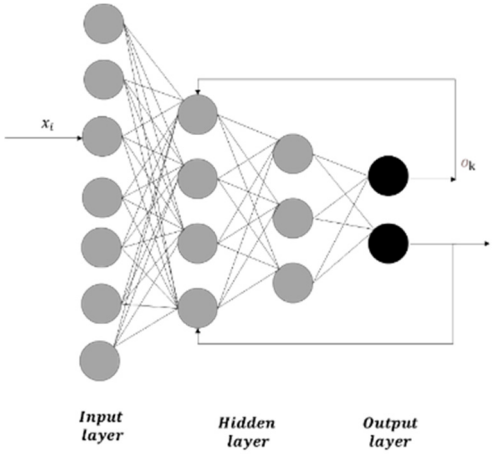

Figure 2. Structure diagram of BP neural network

BP neural network prediction first needs network training, and the training process is as follows.

(1) Signal forward propagation

The output of the input layer is equal to the input signal of the whole network:

$$v_M^m(n) = x(n) \tag{3}$$

The input of i-th neuron such as hidden layer is equal to the weighted sum of $v_M^m(n)$:

$$u_I^i(n) = \sum_{m=1}^{M} \omega_m(n) v_M^m(n) \tag{4}$$

Suppose $f(\bullet)$ is a sigmoid function,

The i-th neuron is equal to the hidden output layer:

$$v_I^i(n) = f(u_I^i(n)) \tag{5}$$

The input of the j-th neuron in the output layer is equal to the weighted sum of $v_I^i(n)$:

$$u_J^j(n) = \sum_{i=1}^{I} \omega_{ij}(n) v_I^i(n) \tag{6}$$

The output of the j-th neuron in the output layer is equal to:

$$v_J^j(n) = g((u_J^j(n))) \tag{7}$$

Error of the j-th neuron in the output layer:

$$e_j(n) = d_j(n) - v_J^j(n) \tag{8}$$

Total error of network:

$$e(n) = \frac{1}{2} \sum_{j=1}^{J} e_j^2(n) \tag{9}$$

ISSN: 2688-9323

(2) Signal back propagation

Firstly, the error between the output layer and the hidden layer is adjusted. In the weight adjustment stage, it is adjusted layer by layer along the network.

Assuming that the number of training samples is single, the error of single sample data is:

$$e_p = \frac{1}{2}\sum_{k=1}^{L}(T_k - o_k)^2 \tag{10}$$

If the number of training samples is p, the error of P sample data is:

$$e_p = \frac{1}{2}\sum_{p=1}^{p}\sum_{k=1}^{L}(T_k^{\ p} - o_k^p)^2 \tag{11}$$

According to the gradient descent method, the connection parameter values in the system can be modified in turn to obtain the expression:

$$\Delta w_{ki} = -\eta \frac{\partial E}{\partial w_{ki}} \tag{12}$$

$$\Delta a_k = -\eta \frac{\partial E}{\partial a_k} \tag{13}$$

$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}} \tag{14}$$

$$\Delta \theta_i = -\eta \frac{\partial E}{\partial \theta_i} \tag{15}$$

The output layer connection parameters can be obtained by combining the four formulas 12, 13, 14 and 15

$$\Delta w_{ki} = -\eta \frac{\partial E}{\partial w_{ki}} = -\eta \frac{\partial E}{net_k}\frac{net_k}{\partial w_{ki}} = -\eta \frac{\partial E}{\partial_{O_k}}\frac{\partial_{O_k}}{\partial net_k} = \frac{\partial net_k}{\partial w_{ki}} \tag{16}$$

The threshold change of the output layer is:

$$\Delta a_k = -\eta \frac{\partial E}{\partial a_k} = -\eta \frac{\partial E}{net_k}\frac{net_k}{\partial a_k} = -\eta \frac{\partial E}{\partial_{O_k}}\frac{\partial_{O_k}}{\partial net_k} = \frac{\partial net_k}{\partial a_k} \tag{17}$$

The weight change of hidden layer is:

$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}} = -\eta \frac{\partial E}{net_k}\frac{net_k}{\partial w_{ij}} = -\eta \frac{\partial E}{\partial_{O_k}}\frac{\partial_{O_k}}{\partial net_k} = \frac{\partial net_k}{\partial w_{ij}} \tag{18}$$

The change threshold of the hidden layer is:

ISSN: 2688-9323

$$\Delta\theta_i = -\eta\frac{\partial E}{\partial\theta_i} = -\eta\frac{\partial E}{net_k}\frac{net_k}{\partial\theta_i} = -\eta\frac{\partial E}{\partial_{O_k}}\frac{\partial_{O_k}}{\partial net_k} = \frac{\partial net_k}{\partial\theta_i} \tag{19}$$

And because:

$$\frac{\partial E}{\partial o_k} = -\sum_{p=1}^{P}\sum_{k=1}^{L}(T_k^p - o_k^p) \tag{20}$$

$$\frac{\partial net_k}{\partial w_{ki}} = y_i; \frac{\partial net_k}{\partial a_k} = 1; \frac{\partial net_k}{\partial w_{ki}} = x_j; \frac{\partial net_k}{\partial\theta_i} = 1 \tag{21}$$

$$\frac{\partial E}{\partial y_i} = -\sum_{p=1}^{P}\sum_{k=1}^{L}(T_k^p - o_k^p)\cdot\psi'(net_k)\cdot w_{ki} \tag{22}$$

$$\frac{\partial y_i}{\partial net_i} = \phi'(net_i) \tag{23}$$

$$\frac{\partial o_k}{\partial net_k} = \psi'(net_k) \tag{24}$$

Finally, the correction quantities of the weight and threshold corresponding to the input layer and output layer of BP neural network are obtained as follows:

$$\Delta w_{ij} = \eta\sum_{p=1}^{P}\sum_{k=1}^{L}(T_k^p - o_k^p)\cdot\psi'(net_k)\cdot w_{ki}\cdot\phi'(net_i)\cdot x_j \tag{25}$$

$$\Delta\theta_i = \eta\sum_{p=1}^{P}\sum_{k=1}^{L}(T_k^p - o_k^p)\cdot\psi'(net_k)\cdot w_{ki}\cdot\phi'(net_i) \tag{26}$$

$$\Delta w_{ki} = \eta\sum_{p=1}^{P}\sum_{k=1}^{L}(T_k^p - o_k^p)\cdot\psi'(net_k)\cdot y_i \tag{27}$$

$$\Delta a_k = \eta\sum_{p=1}^{P}\sum_{k=1}^{L}(T_k^p - o_k^p)\cdot\psi'(net_k) \tag{28}$$

3. Experiment

3.1. Data Cleaning and Preprocessing

After obtaining the original data, the data has some problems, such as poor readability, serious lack of housing information and so on. Data preprocessing is very important in machine learning [9]. So first, clean and transform the data. Data missing value processing mainly adopts mode interpolation for category variables, and completes the replacement processing for missing information, while the missing value and abnormal value of numerical variables are interpolated with their mean value. For the houses with serious lack of information, the whole data was deleted. After cleaning, a total of 1836 houses remained.

ISSN: 2688-9323

According to the correlation analysis, as shown in Figure 3, this paper selects 10 influencing factors with the greatest correlation, which are:Unit price, Area, Floor, Fitment, Type, Type, Equipped with elevator, Tenure of property, Structure, Trading permissions, Region. After analysis and sorting, we have digitized the index. Firstly, the unit of measure of price, unit price, area and ten of property data are removed, and then the remaining indicators are converted. The interpretation of the conversion indicators is as follows:

(1) Floor: Because the total floors of different houses are different, we do not take the floor where the house is located as the division standard of floor variables, but divide the floor variables according to the proportion of floors in the total floor, and divide them into three types: high-rise floor, medium floor and low floor, which are respectively represented by: 3, 2 and 1.

(2) Fitment: According to the crawled data, the decoration is divided into four situations: blank, paperback, hardcover and others, which are represented by 1, 2, 3 and 4 respectively.

(3) Type: The crawled data is m room and n hall, so it is numerically converted to mn, for example, three rooms and two halls: 32.

(4) Equipped with elevator: Distribution elevator is 1, otherwise it is 0.

(5) Structure: It is mainly divided into seven structures: brick concrete structure, brick wood structure, steel-concrete structure, steel structure, mixed structure, frame structure and unknown structure, which are represented by 1, 2, 3, 4, 5, 6 and 7 respectively.

(6) Trading permissions: It is divided into five types: demolition and reconstruction housing, housing reform housing, fund-raising housing, affordable housing and commercial housing, which are represented by 1, 2, 3, 4 and 5 respectively.

(7) Region: This paper mainly obtains the Second-hand housing data of 13 areas in Chongqing. According to the initial Pinyin order, they are Banan District, Beibei District, Bishan District, Dadukou District, Hechuan District, Jiangbei District, Jiangjin District, Jiulongpo District, Nan'an District, Shapingba District, Yubei District, Yuzhong District and Changshou District, which are represented by 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 and 13 in turn.

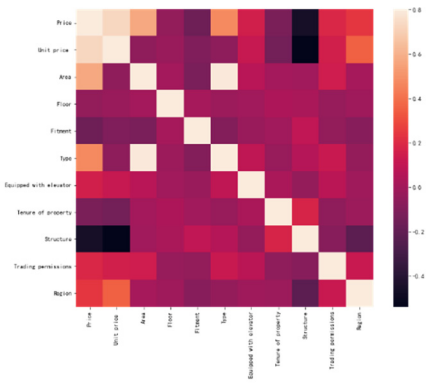

Figure 3. Correlation analysis of factors affecting Second-hand house price

3.2. Normalization Processing

The data collected for modeling often has the problem of dimension, so it is necessary to normalize the original data to eliminate the influence of dimension before formally establishing the model, which is an indispensable basic work. For example, for the information of two variables: house area and floor, the value of area ranges from single digits to hundreds. Relatively speaking, the value of floor is concentrated in the range of 100. This dimensional gap will seriously affect the selection of factors and the analysis of the importance of final influencing factors. Data normalization is to erase the dimensional influence between variables through mathematical processing, so that all influencing factors are on the same line, and

ISSN: 2688-9323

eliminate the interpretation deviation caused by excessive dependence on large dimensional variables, so that the data can be interpreted with more excellent models.

The normalization method used in this paper is as follows:

Z-score Standardization:

$$x^* = \frac{x - \mu}{\sigma} \tag{29}$$

Where, is the mean value of all sample data and the standard deviation of all sample data.

3.3. Evaluation Index

In this paper, the mean square error (MSE), mean absolute error (MAE) and R-squared are used to evaluate the accuracy of the model.

1. MSE:

$$MSE = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \tag{30}$$

2. MAE:

$$MAE = \frac{1}{\mathrm{n}} \sum_{1}^{n} \left| \hat{y}_i - y_i \right| \tag{31}$$

3. R-Squared:

$$R^2 = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2} \tag{32}$$

3.4. Experimental Results and Discussion

Through the above series of experiments, this paper constructs Lasso Second-hand house price prediction model and BP neural network Second-hand house price prediction model respectively, and forecasts the Second-hand house price in Chongqing. The mean absolute error (MAE), mean square error (MSE) and R-squared are calculated by formulas 30, 31 and 32. These three indexes are used to measure the prediction accuracy and prediction performance of each model.

The detailed prediction accuracy data of the two models are shown in Table 1. The mean absolute error of map e is 0.011, and the mean absolute error of map e is 0.085. This shows that compared with Lasso model, BP neural network model has higher prediction accuracy.

Table 1. Prediction error values corresponding to different models

	MAE	MSE	R-Squared
Lasso	0.023	0.116	0.924
BP neural network	0.011	0.089	0.935

In order to more accurately measure the prediction performance of the two models, this paper selects certain test point data in the test set, and draws the predicted value and real data value curves of the Second-hand house prices of the two models in the same figure for comparative analysis of the differences. It can be seen from Figure 4 and figure 5 that the above two models can reflect the overall trend of precipitation data change. By comparing and observing the two models, it can be seen that the predicted value of BP neural network model is consistent with

ISSN: 2688-9323

the real value, and the fitting effect is good, followed by Lasso model. Therefore, BP neural network model can better reflect the fluctuation trend of Second-hand housing price in Chongqing, and is more suitable for the prediction of Second-hand housing price.

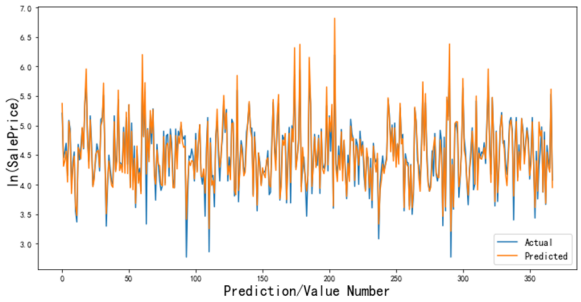

Figure 4. Losso model fitting effect diagram

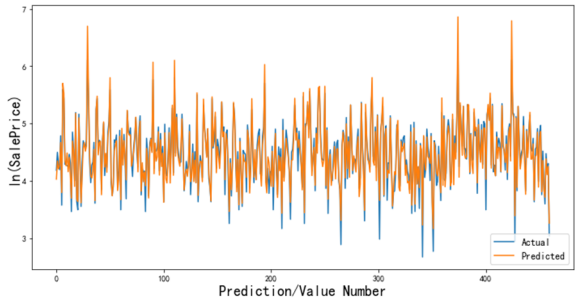

Figure 5. BP neural network model fitting effect diagram

4. Conclusion

This paper uses Python as the main analysis tool, uses the relevant data of Second-hand housing transactions in Chongqing on Lianjia website to study, adheres to the principle of "specific analysis of specific problems", analyzes and processes the data, and constructs Lasso model and BP neural network model for predicting Second-hand housing prices, A variety of model performance evaluation indexes are selected to compare and evaluate the two models. After comparative analysis, the prediction accuracy of BP neural network model as a whole will be better than Lasso model, which is more suitable for the prediction of Second-hand house prices.

Because the house price of Second-hand houses is affected by many random factors, this paper only considers the hard indicators of houses in the prediction of the house price of Second-hand houses in Chongqing, without considering the soft indicators, such as the browsing volume and evaluation of houses, so the predicted value can only be used as a reference. In practical application, we should also combine hard indicators with soft indicators, and combine macro factors with micro factors to make more accurate prediction, so as to provide guidance to buyers, sellers, real estate agents and relevant government regulatory departments of Second-hand housing transactions.

Generally speaking, using mathematical theory and computer technology to predict the price of Second-hand housing is a very meaningful research direction.

References

[1] Shi Xiaohao, Gao Juan, Kan Xiaojing Population mobility shift: from one-way urbanization to weak urban-rural integration -- a comprehensive analysis based on multi-source data in Shandong Province [J] Economic trends and reviews, 2020 (02): 27-42 + 218.

ISSN: 2688-9323

[2] Dong Tian The volume and price rise are stable, and the Second-hand housing market in Beijing has obviously warmed up [n] China Securities Journal, 2022-01-19 (A07) DOI:10.28162/n. CNKI. nczjb. 2022.000329.

[3] Ren Shijie Research on real estate value evaluation method along urban rail transit based on BP neural network Beijing Jiaotong University, 2015.

[4] Wang Jingxing Research on house price prediction model based on regression [J] National circulation economy, 2020 (19): 3.

[5] Embaye WT, Zereyesus YA, Chen B (2021) Predicting the rental value of houses in household surveys in Tanzania, Uganda and Malawi: Evaluations of hedonic pricing and machine learning approaches. PLoS ONE 16(2): e0244953.

[6] Wang Lu, sun Ju Bo Application of Lasso regression method in characteristic variable selection [J] Journal of Jilin Normal University of engineering and technology, 2021,37 (12): 109-112.

[7] Zhang Hongbin, Guo Meng Machine learning and financial forecasting -- An Empirical Study on the application of performance mine warning in Chinese Listed Companies [J] Quarterly Journal of finance, 2020,14 (04): 135-154.

[8] Xie Hao Vehicle speed prediction based on BP neural network and its optimization algorithm [D] Chongqing University, 2014.

[9] Tang Chenglong, Chen Wei, Tang Haichun, Wu Zefeng Research and application of data preprocessing methods in the context of big data [J] Information recording materials, 2021,22 (09): 199-200 DOI:10.16009/j.cnki. cn13-1295/tq. 2021.09.094.

[10] Yang Dashan, Liu Wei Study on Influencing Factors of Second-hand housing transaction price based on Hedonic Price Model -- Taking Shanghai Second-hand housing transaction market as an example [J] Modern business, 2020 (30): 42-45 DOI:10.14097/j. CNKI. 5392/2020.30.017.