Russian housing price forecast based on BP neural Network

Sidi Du[1, †], Xiqing Liu[2, †], Songze Shi[3, *, †] and Guanchen Zhou[4, †]

[1]School of Economics, Guizhou University, Guiyang, China

[2]School of international programs, Guangdong University of Finance, Guangzhou, China

[3]Department of Math, Beijing Normal University, Beijing, China

[4]Malvern College Chengdu, Chengdu, China

*Corresponding author: 201911130118@mail.bnu.edu.cn

[†]These authors contributed equally.

Abstract. Contemporarily, Russia's real estate market has huge potential for investors. The purpose of this paper is to investigated the impact of various factors on the real estate market price based on various statistical models. We collected past housing price data from the Internet. Subsequently, PCA model was adopted to reduce the dimension of the data to simplify the calculation. Finally, the BP model is implemented to predict the price changes in the real estate market. After calculation, our forecast data and the reality of housing prices are slightly different, but the forecast of the real estate market is consistent with the reality. In order to further find the influence of various factors on housing price, we also calculated the characteristic importance of different factors and found the most important influencing factor, i.e., the region. These results can provide investors with an idea to analyze the real estate market prices, and help policy makers find the real estate market prices in line with the market rules to avoid the occurrence of bubbles.

Keywords: PCA model; BP model; feature importance; price.

1. Introduction

High level of real estate development is one of the necessary conditions for the effective operation of national economy. In Russia, similar to other countries, the real estate market has become one of the most capital-intensive regions [1]. The global financial crisis of 2008 limited Russia's economic momentum. With the spread of the European debt crisis and the outbreak of the Ukraine crisis in 2014, the Russian economy gradually fell into trouble [2]. In addition, since 2020, due to the impact of COVID-19, Russian GDP growth rate has dropped to a new low. Besides, the inflation rate caused by the Russia-Ukraine conflict has once again caused a strong impact on Russia's economy and financial system. Therefore, whether real estate market affected in the face of economic downturn and what factors affect the house prices are remained unclear. In fact, though the Moscow real estate market has shown a weak state, there has been no housing price collapse. In contrast, there has been little fluctuation in the eight years from 2008 to 2016, and even maintained an average annual growth rate of 7% [2].

Besides, the internal macro-control and micro-economic factors also have an impact on the price of Russian real estate. We have searched the housing price in Russia from 2018 to 2021 and find a graph that describe the change of price. As a matter of fact, the price of Russian housing price has fallen at first and then there were several fluctuations during 2020. After that the price raises continuously, the estate market is sensible and the factors of price of Russian estate which are analyzed in this paper include latitude, longitude, region of Russia, apartment floor, number of storey, the number of living room. In this paper, we will use the relationship between these factors and the price of Russian housing price to predict the change of the price in the future.

This paper uses machine learning to predict housing prices in order to have a better understanding of the estate market. Compared with other methods, machine learning is more capable of learning, and it can process more data than human labor. In addition, machine learning has obvious advantages in nonlinear and multivariate statistics, which shows a good application prospect in housing price

prediction. More importantly, there are a few empirical studies using ML in housing price forecast. Besides, we use Correlation Analysis, Principal Component Analysis and Back Propagation Neural Network to find the change of the housing price in Russia in the future. In this paper, by using machine learning method to forecast the price at the micro level, looking forward to helping house sellers and buyers to make more scientific decisions, and bring certain inspiration to the subsequent research of machine learning in terms of housing forecast. Moreover, having a clear understanding of the effects of the factors on estate price can help the government adjust real estate market in first-tier cities in order to prevent the financial crisis.

In regard to the exist studies and other methodological and theoretical basis of the paper, many of them have conducted in-depth analysis and research on the influencing factors of house prices. In perspective of macroeconomics. Xu and Peng studied and compared the determinants of house prices in China, the United States, Japan, the United Kingdom, Russia and India from the perspectives of economic growth, price levels and monetary policy. According to the results, the main factors affecting house prices in developed countries and developing countries are different. The macroeconomic sensitivity of the real estate market in developed countries is significant, while the real estate market in developing countries is affected by institutional factors, rigid demand, traditional habits and local environment [3].

In perspective of micro level, Jin found that the characteristic variables that affect the price of commercial housing in Sanya include 8 variables: Decoration degree, plot ratio, property management, parking space, floor where the house is located, building area, house type and living facilities [4]. Some scholars have also conducted specific research on the Russian real estate market. Zhang judged, in 2015 ~ 2016, the real estate market of Moscow city should still be a stagnant and declining market [5-8]. That means lower demand growth, new housing development and delivery, housing supply, affordability and prices. It is estimated that the demand for housing in the primary market may drop by 15% to 20% in 2015, resulting in oversupply. Housing uptake in the market could fall by 10% and prices could continue to fall. Secondary market demand may show a slight decline in 2015, and may decline by 3% ~ 4% in 2016 [4, 9, 10]. Ren studied the situation of housing prices in Moscow from 2009 to 2016 from 6 factors: benchmark interest rate, M2, the exchange rate of the US dollar ruble, Moscow population, income level, the area of new housing in Moscow and other factors. His research shows that the main factors affecting housing prices in Moscow are the exchange rate, the area of new housing and population, and predicts that the real estate market in Moscow will gradually recover after 2017 [2].

To sum up, the previous studies on the influencing factors of house prices are relatively rich, but there are few studies based on the micro environment of the Russian real estate market.

2. Data & Method

2.1 Data

We collect the data from the Kaggle website of datasets case from 2018 to 2021. The data shows the price of Russian house which is the basis of this paper. There is a massive number of data included in this DS. Therefore, in order to simplify calculation, we choose one of the data from each day. The data also include ten factors which are latitude, longitude, region of Russia, apartment floor, number of stores, and the number of living room [5].

2.2 PCA model

Principal component analysis (PCA) is a common multivariate statistical method. Since its dimensionality reduction idea is very close to the requirement of multi-index evaluation index sequence, it has been applied to the evaluation of sociology, economics and management in recent years and gradually become a unique multi-index evaluation technology. However, there are many problems in applying principal component analysis directly to evaluation. This is mainly because

statistical analysis, as a mathematical means from the outside to the inside, emphasizes its objectivity, while evaluation theory's view of objective things is based on the appraiser's value judgment [6].

PCA is a data analysis method proposed by K. Pearson more than a century ago. Its starting point is to calculate a group of new features in order of importance from a group of features. They are linear combinations of the original features and are unrelated to each other. Principal component analysis is suitable for demography, quantitative geography and so on. In short, one of the methods of statistical analysis is principal component analysis that divides many variables into a few comprehensive indicators. The calculation steps of PCA are as follows.

Assuming that there are n samples and P indexes (variables), the sample matrix x of N×P can be formed.

- Step 1: Standardize the sample matrix X
  Average by column and the standard deviation Calculate standardized data , the original matrix is normalized to:

- Step 2: It is to calculate the covariance matrix of X

Here,

It should be noted that the above two steps can be directly combined into one step: directly calculate the correlation coefficient matrix of x matrix:

The formula expresses that is equal to the correlation coefficient between the first column of x and the second column, that is, the correlation coefficient between variable 1 and variable 2.

- Step 3: It is to calculate the eigenvalues and eigenvectors of R. As for Characteristics: (is a semidefinite matrix, and), the Feature vector will be

- step 4: It is to calculate the principal component contribution rate and cumulative contribution rate

- Step 5: Write the principal components. Generally, the first, second, and... corresponding to the characteristic values whose cumulative contribution rate exceeds 80% are taken and the m(m≤p) principal component. The ith is in the principal component:

2.3 Introduction of BP neural network

BP neural network is the learning process of error back propagation algorithm, which is composed of two procedure: forward propagation of information and back propagation of error. This network has three components: input layer, hidden layer and output layer. The input layer is responsible for receiving input information from the outside and passing it to the hidden layer; The hidden layer is the internal information processing layer, which is responsible for data processing and information transformation. According to the requirements of information change ability, it can be designed as a single hidden layer or a multi-hidden layer structure. Finally, the hidden layer further processes the information and outputs the results through the output layer, which is a forward propagation process of learning and training.

In BP neural network, the back-propagation stage of error is also of vital importance. Through this procedure, one can constantly adjust the weight of each layer, i.e., the established network can more precisely approach the real data. When the final output data is inconsistent with the real data, it enters the back-propagation stage of error. The value of the error is represented by the loss function:

i.e., the absolute value of the difference between the output value (recorded as) and the real value (recorded as), which can directly reflect the value of the result error. In order to improve the accuracy of the model, the weight of parameters must be adjusted to reduce the value of the loss function.

Using the error gradient descent method can correct the weight of each layer and finish back propagation layer by layer. Hence, the process of back propagation can be understood as calculating the gradient (,) .... etc.) of each parameter (such as,... etc.) back according to loss, and then add their corresponding gradients to the original parameters. So, this repeated process of information forward propagation and error back propagation adjusts the weight of each layer of parameters, and then minimizing the overall error in this learning and training process. In this way, it can have more accurate predictions and achieve a better model effect. Over all, this is a general introduction about how BP neural network operates.

As the linear function has no upper bound, the data may be too high and the calculation work can be enormous for computer to process. By using Sigmoid function, the data can be limited to a certain range. In addition, the image of sigmoid function is a nonlinear curve, so it can better approximate the nonlinear relationship, because in most cases, the actual relationship is nonlinear. The Sigmoid function is called the excitation function, and its mathematical expression is:

Its derivative with respect to X can be expressed as:

In terms of data processing, in order to avoid defects such as local error optimization or error oscillation caused by too large discrete or too centralized absolute values of sample parameters, the data must be normalized, that is, all data must be compressed within the range of (0, 1). Moreover, normalization can remove the influence of the dimensional magnitude of each parameter on the data, giving equal importance to the input variables with different physical meanings. It can greatly avoid neuron saturation and improve the convergence of the network. The formula of normalization is as follows:

3. Results and Discussion

3.1 PCA

The influencing factors of house prices are complex. Based on the previous research results and combined with database of Russia, this paper takes the sales price of commercial housing as the output variable y, with 10 input variable x :the geography latitude (), the geography longitude (), region (), building type (), level (), levels (), rooms (), area (), kitchen area () and object type (). MATLAB analysis software is used to calculate the relationship between the economic indicators of the selected ten main influencing factors and the price of Russia commercial housing. The correlation coefficient matrix of the obtained samples is shown in Table 1. After calculation, the eigenvalues, corresponding eigenvectors and contribution rates of the correlation coefficient matrix are listed in the following Table 2.

Table. 1 The correlation coefficient matrix (R)。

	1.00	-0.01	-0.08	-0.13	0.05	0.11	-0.04	-0.04	-0.03	0.03
	-0.01	1.00	0.62	-0.10	-0.12	-0.12	0.03	-0.03	-0.06	-0.02
	-0.08	0.62	1.00	-0.02	-0.12	-0.15	0.05	0.00	-0.03	0.03
	-0.13	-0.10	-0.02	1.00	-0.08	-0.21	0.04	0.05	0.05	-0.07
	0.05	-0.12	-0.12	-0.08	1.00	0.65	-0.05	0.07	0.18	0.17
	0.11	-0.12	-0.15	-0.21	0.65	1.00	-0.14	0.05	0.27	0.32
	-0.04	0.03	0.05	0.04	-0.05	-0.14	1.00	0.75	0.28	-0.13
	-0.04	-0.03	0.00	0.05	0.07	0.05	0.75	1.00	0.57	0.00
	-0.03	-0.06	-0.03	0.05	0.18	0.27	0.28	0.57	1.00	0.17
	0.03	-0.02	0.03	-0.07	0.17	0.32	-0.13	0.00	0.17	1.00

Table. 2 the eigenvalues, corresponding eigenvectors and contribution rates of R

Eigenvector										
: geography longitude	0.04	-0.15	-0.05	-0.70	-0.56	-0.40	0.01	0.07	0.04	0.00
: geography latitude	-0.22	0.22	-0.62	-0.01	0.04	-0.17	0.08	-0.69	0.08	0.01
:region	-0.20	0.26	-0.60	0.14	-0.02	-0.14	-0.03	0.69	-0.12	0.01
:building type	-0.04	0.17	0.29	0.57	-0.38	-0.62	-0.11	-0.10	-0.14	-0.01
:level	0.39	-0.33	-0.17	0.10	0.35	-0.41	-0.26	0.09	0.58	0.00
:levels	0.42	-0.40	-0.24	0.03	0.15	-0.13	0.03	-0.09	-0.74	-0.04
:rooms	0.30	0.52	0.05	-0.21	0.10	0.02	-0.47	-0.04	-0.11	-0.59
:area	0.46	0.45	0.01	-0.10	-0.01	0.04	-0.09	-0.02	-0.04	0.75
:kitchen area	0.49	0.19	-0.08	0.15	-0.19	0.05	0.73	0.05	0.20	-0.30
:object type	0.20	-0.23	-0.28	0.29	-0.59	0.47	-0.39	-0.07	0.13	-0.02
characteristic value	0.18	0.29	0.37	0.59	0.83	0.89	1.08	1.56	2.02	2.20
Contribution	0.22	0.20	0.16	0.11	0.09	0.08	0.06	0.04	0.03	0.02
Cumulative contribution rate	0.22	0.42	0.58	0.69	0.77	0.86	0.92	0.95	0.98	1.00

Generally, when the cumulative contribution rate is greater than 85%, it is considered to be sufficient to reflect the original variable information, and the corresponding m is the first m principal components taken. It can be seen from the table that the cumulative contribution rate of the first six principal components reaches 86%. Therefore, we select the first six principal components, which can well summarize the original variables.

To analysis of load of principal component, following results are presented

Here, is the standardized coefficient, : the geography latitude, the geography longitude, region, building type, level, levels, rooms, area, kitchen area and object type. The first principal component has a moderate positive load on,, , , . Therefore, the first principal component can be explained as the comprehensive reflection of buildings internal structural on housing price. The second principal component has a large positive load on and , and a moderately similar positive load on. Therefore, can be explained as comparing the impact of interior space(rooms number and area) with location (region and geography latitude), reflecting an important influence of interior space on Russian house price. The third principal component has a moderate positive load on , while a large negative load on , and approximately equal negative load or very small positive load on another component. Therefore, can be explained as house price is negatively correlated with location (geography latitude and region) while positively with the building type. The fourth principal component has a large positive load on , while a large negative load on . Therefore, can be explained as house price is negatively correlated with location (geography longitude) while positively with the building type. The fifth principal component has a moderate positive load on , and approximately equal negative load or very small positive load on other component. Therefore, calling as the level component. The sixth principal component has a moderate positive load on as well as approximately equal negative load or very small positive load on other component. Therefore, calling as the object type component.

From the results reflected by these six principal components, comparing with the external environment of the houses, the internal conditions of them have a greater impact on the house pricing. As a result, the internal factors of the house should be fully considered in the process of predicting real estate price.

3.2 BP model

Through the neural network analysis of MATLAB, 1045 groups of data were divided into three different sets: the training set, the verification set and the test set. And their proportion is 75%, 15% and 15% respectively. The input data is six main factors affecting Russian house prices, and the output is Russian house prices. Ten hidden layer neurons are set in this model, and the training method is Levenberg marquardtal algorithm. The specific parameters of the model are shown in the Table. 3. MSE(MeanSquaredError) reflects the error between the output data and the real data of the test set, and its expression is , where is the output data, is the real data). The following figure shows the fluctuation trend of mean square error with the number of training. As shown below, the vertical axis prsents the value of MSE (means quare error) and the horizontal axis presents the increasing progress of epochs.

Generally speaking, after many times of training stages, the error will decrease, but as the network begins to over fit the training data, the error of the validation data set may begin to increase. It can be seen from the image that after the MSE of the validation data set increases six times continuously, the training stops, and the best model corresponds to the smallest MSE. R^2（GoodnessofFit）indicates the extent to which the model explains the original data. Its value is between (0, 1). The more closer the value to 1, the smaller the error of the model is and the better the fitting effect is. The formulae of R^2 is expressed as:

After a series of data processing, the results of the model in training set, verification set and test set are =0.90436, =0.85295, =0.82608, =0.88021. The image shows that the model performs best in the training set, reaching 0.90436, followed by the verification set, and reached only 0.82608 in the test set. However, the overall value of the model is still 0.88021, which shows that the model has good explanatory ability in general, and has a relatively accurate prediction of Russian house prices, which can be used as an effective method to predict house prices.

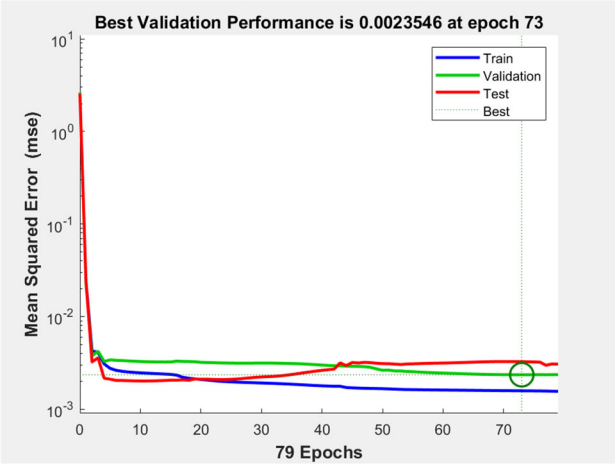

Fig. 1. Changes of MSE in 79 epochs.

Table. 3 The specific parameters of the model

The number of input factors	The number of hidden layers	The number of epoch progress
6	10	79

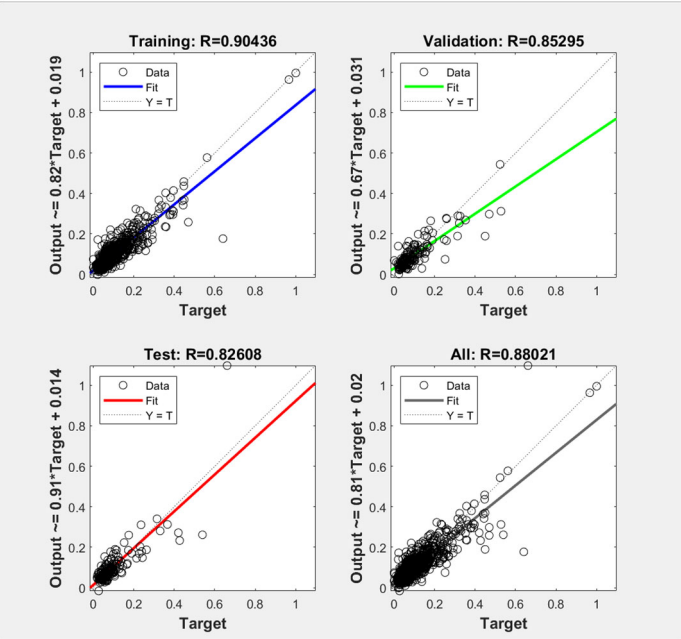

Fig. 2. R2 in different conditions.

According to the results, the model performs best in the training set, reaching 0.90436, followed by the verification set, and reached only 0.82608 in the test set. However, the overall value of the model is still 0.88021, which shows that the model has good explanatory ability in general, and has a relatively accurate prediction of Russian house prices, which can be used as an effective method to predict house prices. In view of the good explanatory ability of the model, we predict the price of Russian houses in 30 days. After predicting the result of the Russian house price, we find the feature importance to find the contribution of different factors. The region and longitude have a great influence on the price. Based on the analysis, the area of house in Russia has the biggest impact on the price and its weight is 0.473298. Whereas the object_type has the smallest impact on the price,and its weight is just 0.002933. Besides, the region and longitude have a great influence on the price.

4. Summary

In summary, this paper investigates the future price of the estate price of Russia based on the data on the kaggle. According to the analysis, the region has a huge impact on the real estate price, so investors should focus on the region where the house is located. Investors can consider house prices comprehensively following the characteristic importance of different factors to avoid losses. For policy makers, comprehensive regulation can be carried out in terms of the impact of different factors on housing price, which can find the real housing price in line with the market and avoid the occurrence of housing price bubble.

Owing to the huge amount of data collected online, up to hundreds of thousands of groups, we selected a group of data from the daily data, which may lose part of the information, leading to certain errors in the prediction results. Since there are only ten influencing factors selected, they cannot all reflect the impact of real-life housing price, leading to the deviation of prediction results. PCA model reflects most of the information, while a small part cannot be reflected by the selected principal components. Therefore, the lack of information collected in BP model results in the deviation of results. In the future, therewill be volatility in the Russian property market. Overall, these results offer a guideline for investors and policy-maker to have a clear understanding of Russian real estate market.

References

[1] Fadeeva Anastasia, State Price Regulation at the Residential Real Estate Market in Russian Federation, Chongqing, Chongqing University, 2018

[2] Ren Jun, Studies on the Influencing Factors of Moscow Housing Price since global financial crisis, Beijing, University of international business and Economics, 2017

[3] Xu Jun, Peng Zhenjiang, International Comparison of the determinants of real estate prices, Hebei Finance, 2011

[4] Zhang Congming. A Review of the Current Situation of Russia's Real estate Market. Eurasian Economics 6(2015):17.

[5] Information on: https://www.kaggle.com/datasets/mrdaniilak/russia-real-estate-20182021

[6] Jinghua Li and Yaohuang Guo. Principal component Analysis for multi-index evaluation: Principal component evaluation. Journal of Management Engineering, 16(1), 5, 2002.

[7] Mohammcd E, Sudhakar K V. ANN Back-propagation Pre- diction Model for Fracture Toughness in Micro Alloy Steel. International Journal of Fatigue, 2002, 24(9): 1003-1010.

[8] Rui Xu, and Fengxian Han. Research progress and application of BP neural network to predict housing price along subway line. Sichuan building materials, 47.12(2021): 5.

[9] Liuhao Chen. Neural network BP algorithm research summary computer knowledge and technology. 2010. 36(6) :103641036.

[10] Yuming Gao, and Renjin Zhang. Housing price forecast analysis based on genetic algorithm and BP neural network. Computer engineering, 2014, 40(4):5.