Research on Housing Price Prediction Based on Machine Learning

Zirui Huang*

China Hangzhou Dipont School of Arts and Sciences, Hangzhou, Zhejiang, China, 311121

* Corresponding Author Email: huang.zirui@rkcshz.cn

Abstract. The stability and long-term growth of the real estate industry depend heavily on the capacity to forecast house price trends with accuracy. The purpose of this research is to evaluate the benefits and drawbacks of several machine learning models for housing price prediction. First, the factors influencing housing prices in various ways are explained. Next, the multiple linear regression, backpropagation neural network, and random forest model, respectively, are introduced. In the meantime, the paper analyzes their forecasting shortcomings and provides examples of other scholars' optimization experiments on the aforementioned three models. The findings indicate that these three types of optimized models can obtain more stable and accurate prediction results when housing prices are predicted. Housing price forecasting can lower investment risks for real estate developers by assisting them in creating more effective project development plans and sales tactics. The government can better grasp the real estate market's development trend, create pertinent regulations and control measures, and support the stable and sustainable growth of the real estate market with the aid of home price prediction. Finding an appropriate way to predict the price of housing is therefore imperative.

Keywords: Housing price forecast, Multiple Linear Regression, Backpropagation neural network, Random Forest.

1. Introduction

The nation's economy and people's means of subsistence are progressively becoming heated topics as a result of the rapid rise in property prices. In addition to having a direct impact on people's living standards, the sharp rise in urban housing costs also introduces several unstable elements into the social and economic development process. As a result, the precise forecasting of home prices has progressively gained attention from academics.

In the use of machine learning models to predict housing prices, Chinese scholars have carried out in-depth research. Zhou Liangjin and Zhao Mingyang conducted a random forest-based prediction analysis of second-hand housing prices in Shenzhen, and the findings demonstrate that the random forest model has significant flexibility, reliable outcomes, and ease of implementation [1]. Song Yaotook the house price data of Boston as the starting point and introduced the random forest regression model and the fully connected regression model respectively to predict other house prices in the city [2]. The final result showed that the fully connected model scheme was superior to the random forest scheme when RMSE was used as the measurement index, indicating that the fully connected scheme was a good scheme for predicting house prices. Zhou Liangjin and Zhao Mingyang built models based on the K-nearest neighbor approach, decision tree, random forest, and support vector machine, respectively, for test prediction [3]. They did this by selecting housing prices and related data from 35 large and medium-sized cities in China from 1998 to 2019. The findings demonstrate that, in terms of fit degree, the K-nearest neighbor method outperforms decision trees, random forests, and support vector machines.

This study compares the benefits and drawbacks of various forecasting techniques, examines the benefits and drawbacks of utilizing various machine learning models to predict housing prices, and provides an analysis of several improvement techniques with examples.

2. Factors Affecting Housing Prices

Correctly collecting the factors that affect house prices from different aspects is crucial when using machine learning models to forecast house prices. By analyzing a large number of data and related features, the key factors affecting the housing price are extracted, and the corresponding prediction model is established, which can better reflect the change rule of the housing price and improve the accuracy of the prediction.

2.1.Economic Development

Cities have grown and modernized as a direct result of economic progress. The expansion of the national economy has also raised demand for homes, which has driven up house prices even further. The real estate market bubble will also be exacerbated by the decline in housing liquidity brought on by economic progress [4]. After more than 20 years of rapid development, the real estate industry once became a "pillar industry" to promote economic development due to its numerous related industries, which not only led to the real estate market and the macroeconomy being tightly tied together but also caused the soaring housing prices and overheating of the real estate economy.

2.2.Regional Factor

The real estate market has unique regional characteristics, and the housing price in different regions is affected by many factors. In general, housing prices tend to be higher in urban centers because these areas have convenient transportation and high-quality education and medical resources, which attract a large influx of people and high demand for housing [5]. Generally speaking, houses have different social, economic, and cultural locations due to their different social, economic, and cultural locations in cities or regions, and their prices in the market are also significantly different. In addition, because the conditions determining the location are developing and changing, housing prices will also change with the change of location. It can be said that location conditions determine the price level of real estate.

2.3.Policy Factor

On the one hand, the government has directly limited the rapid rise of housing prices through the implementation of purchase restrictions, loan restrictions, and other policies. For example, limiting the ability of some buyers to buy a house, reducing market demand, increasing the cost of buying a house, etc., thereby slowing down the pace of price growth. On the other hand, the government influences the demand of the real estate market through policies such as land supply. If the supply of land is sufficient, the government will often encourage the development of housing leasing, and the upward pressure on housing prices will be eased to a certain extent [6]. The research results show that the housing price in Beijing presents a regular rise and fall with the change in the policy regulation cycle. In the policy encouragement period, the housing price shows an overall upward trend, while in the policy contraction period, the housing price shows a downward trend, indicating that the housing price is sensitive to the policy regulation. It is also found that the policy has a certain lag in the regulation of housing prices, and the lag period is about 3 months.

2.4.Housing Supply Structure

To begin with, the price of a home will be directly impacted by how sensible the housing supply system is. An unequal increase in housing prices will result from an illogical housing supply structure, such as an abundant supply of homes in some locations and a deficiency of homes in other places. Thus, one of the key strategies for containing the price of housing is to preserve the rationality of the housing supply structure. Second, home costs will be impacted by the variety of housing supply structures. Housing prices will fluctuate if the housing supply structure is overly simplistic, consisting solely of affordable or commercial housing. Thus, preserving the diversity of the housing supply structure is also a crucial control strategy [7]. It can be seen from the data table of new commodity

housing supply and price from 1995 to 2005 that both housing supply and housing price show an increasing trend, and there is a positive correlation between the two.

3. Machine Learning-Based House Price Prediction

3.1.Multiple Linear Regression

Multiple linear regression is a traditional statistical method usually used to analyze the linear relationship between two or more predictor variables and response variables. Let y be the dependent variable and x_1, x_2 and $x_n$be the independent variable, then the model formula of multiple linear regression is as follows: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon$. In the formula, β_0 , β_1 , $\beta_2$and $\beta_n$are parameters to be estimated, called regression coefficient, and ε is a random variable.

The purpose of using this method is to try to find a system of linear equations in which the error between the predicted values and the actual observed values is minimized.

The stages for creating a regression model are as follows: First, gather a collection of data with independent and dependent variables. The outcomes that are affected by independent factors, like home prices, are the dependent variables. After that, preprocessing is done on the gathered data, which includes sorting, eliminating outliers, adding missing values, changing data types, etc. Next, the best regression model—multiple linear regression or basic linear regression, for example—is chosen based on the type of problem. Then, using least squares or other optimization techniques, the model's parameters are computed. Next, the model is put to the test using residual analysis and other techniques to determine its validity and reliability. Ultimately, the trained model is applied to forecast and examine fresh data, assess the model's impact, and provide a visual report or chart that illustrates the modeling procedure and outcomes. Li Shengda used the Manhattan home prices from 2010 to 2016 as a data set to forecast the 2017 local home prices [8]. The experimental findings show an error rate of approximately 2%, falling within the small probability range. This indicates that the multiple linear regression model is capable of accurately predicting the price of housing in Manhattan.

3.2.Back-Propagation Neural Network

As a type of learning technique for multi-layer artificial neural networks, back-propagation neural networks primarily use the error back-propagation algorithm to continuously modify the weight and bias to reduce network error and ultimately provide the desired output result. The input layer, hidden layer, and output layer are the three layers of a backpropagation neural network. To put it simply, the formula y=ax+b can be applied. The input layer, often known as the neuron, is what relates to x (the current data). The weight and bias, or ab, are represented by the concealed layer. The output layer (the price to predict) is the outcome of fine-tuning the weight bias operation.

The back-propagation neural network uses the gradient descent technique and continuously modifies the network's threshold and weight using backpropagation to reduce the mean square error between the network's expected and actual output values. Both input signal forward propagation and error signal backpropagation are used in the training process. This cycle of computing the output error and modifying the weight continues until the mean square error approaches the predetermined level. The following is an introduction to the specific steps:

Building a neural network requires preparing matching tests and training data sets, such as those including information on the variables influencing home values. The model's parameters are then set at random, including the input, hidden, and output layers' weights and biases. During the training phase, the gradient descent method is used to update the parameters after the forward propagation, loss function, and error backpropagation (to derive the gradient) calculations are completed in order. On the training set of data, the model can get high accuracy or fast convergence by iteratively tweaking the model parameters. Wang Jing, Luo Weiping, and Chen Yongheng forecasted Shanghai's housing prices from 2001 to 2003 using a neural network model and data on the prices of homes in a specific neighborhood of Shanghai provided by Lianjia Company [9]. The findings demonstrate that

the relative error will rise as the year goes on, which highlights the shortcomings of the single variable prediction model.

3.3.Random Forest

Random forest algorithm is an ensemble learning method used to solve classification and regression problems. It is an integrated model composed of multiple decision trees. Each decision tree is trained independently to make a final prediction by voting or averaging. By introducing randomness, the generalization ability and robustness of the model are improved. Make it perform well in a variety of data situations.

A decision tree is a prediction model based on feature selection. Initially, a portion of the original features is chosen at random to serve as the foundation for the decision tree. Principal component analysis, correlation analysis, and other approaches are commonly used for feature screening. From the chosen sample data and attributes, several decision trees are then constructed (each tree is trained using a subset). Ultimately, by combining the predictions from several decision trees, the ultimate prediction outcome is produced. To further enhance the model's performance, pertinent parameters—such as the tree's depth and number of leaf nodes—can be changed based on the model's performance. Li Hanyu, Wei Jiayin, and Lu Youjun selected second-hand real estate data of various dimensions in Shenzhen between March and July 2020, trained the real estate price model using the random forest algorithm with cross-verification, and finally evaluated the model using mean square error and goodness of fit [10]. The findings demonstrate the prediction model's high stability and accuracy.

4. Advantages

4.1.Multiple Linear Regression

Stability and comprehensiveness are two crucial aspects of multiple regression linear models. Comprehensiveness refers to the ability to simultaneously analyze the impacts of several independent variables on dependent variables to fully comprehend the link between the data. This is crucial since numerous variables interact to generate a lot of issues in real life. Stability refers to the multiple regression linear model's relatively consistent prediction accuracy, making it appropriate for long-term forecasting.

4.2.Back-propagation Neural Network

Strong fault tolerance and parallel processing capabilities are features of neural networks. Neural networks are capable of handling vast volumes of data with efficiency since each neuron can compute independently. The neural network also has a high fault tolerance, so even if a few neurons are destroyed, the system's functionality won't be significantly impacted. Because of this, the neural network exhibits excellent stability and robustness when handling challenging tasks.

4.3.Random Forest

As it constructs each decision tree, a random forest takes random samples of the training data. This implies that distinct subsets of the data are used to train each tree, and this variety enhances the model's capacity for generalization. The random forest only considers partially random attributes in each split of each decision tree. Because each split is not dependent on a small number of variables because of this randomization, the model is more robust and broadly applicable.

5. Improvement

5.1.Multiple Linear Regression

When utilizing multiple linear regression, the issue of multicollinearity frequently arises. When two independent variables in a regression study have a high linear connection, this is referred to as

multicollinearity. The model's predicted parameters may not be appropriately estimated when there is multicollinearity among the independent variables, which reduces the model's predictive power. Using the variance inflation factor is the most popular method for making this correction. Wang Jinyi, Hong Zhiyong, and Luo Bowei talk about utilizing the multiple linear regression model to forecast home prices [11]. The data set utilized was the Boston area's home prices for 2019. The variance inflation factor was used, and outliers that were too high or low for each characteristic were removed in order to potentially reduce multicollinearity in the model. The outcomes demonstrated that the optimized multiple linear regression model's prediction accuracy had increased and that the model's accuracy and capacity for generalization had further improved.

Multicollinearity can be quantified using the Variance Inflation Factor (VIF). It displays how well independent variables are correlated. The formula used to calculate is $VIF_i = \frac{1}{1-R_i^2}$. Based on the evaluation of the VIF value, we may select the suitable model and parameter modification technique in a practical setting. Generally speaking, features with excessively high VIF values (more than 10) have substantial multicollinearity relationships with other features. Once the issue has been identified, the model can be improved to enhance the regression model's predictive performance.

5.2.Back-propagation Neural Network（BP neural network）

To minimize the loss function, the network weights and bias parameters are adjusted continuously during the model training process. The BP neural network will use gradient descent to update the weight and bias during this phase, causing the loss function to progressively converge to the minimal value. However, Genetic Algorithm (GA) can be employed to improve the model's global search capability because the gradient descent method's iterative process may lead to the local optimal solution rather than the global optimal solution. The GeneticAlgorithm (GA) is an optimization search algorithm designed to mimic the natural evolution of organisms. It iteratively searches a group of individuals and increases the fitness of individuals in the population to discover the best solution by mimicking the mechanisms of cross-variation and natural selection.

Bu Yujia conducted a comparison between the BP and GA-BP neural networks, using training data from Chengdu's 2006–2007 housing price and related parameters, and utilized the BP and GA–BP neural networks to forecast housing prices in 2018 and 2020, respectively [12]. The findings indicate that the BP neural network is vulnerable to local optimization issues and that the original model's convergence speed and prediction accuracy can be enhanced by the genetic algorithm's strong global search capability.

First, each neuron is assigned a fitness function, which is typically the mean square error (MSE) or another performance parameter. The first set of chromosomes, or weight and bias combinations, are then created at random and used as the genetic algorithm's population. After that, each person's fitness value is assessed, a selection process is used in accordance with the fitness value, and some people are chosen to be the parents of the population that will follow them. To create new kid persons, cross-operation is carried out between parent individuals. Good genes are to be dispersed across the population. Subsequently, the progeny individuals undergo genetic mutations to enhance the search space's diversity. Ultimately, the freshly created progeny individual's fitness value is assessed, and the selection, crossover, and mutation processes are iteratively optimized according to the fitness value. In this sense, the BP neural network's convergence speed and prediction accuracy can be further enhanced by the genetic algorithm.

5.3.Random Forest

The primary challenges with adopting random forest models are non-equilibrium data and continuous variables; thus, the researchers include data pretreatment within the optimization category. The method performs better after preparing the data. Furthermore, improvements have been made to certain data types and domains that would not be able to be managed otherwise. Wu Qiong et al. proposed integrating Neighborhood Cleaning Rule (NCL) technology with the Random Forest

method [13]. NCL technology was primarily used to process the data in the unbalanced training set, and the Random Forest technique was then used to classify the processed data.

Non-equilibrium data set refers to the fact that the number of one type of sample in the data set is much more than or less than the number of other samples, and the use of a conventional random forest algorithm on this premise will lead to uneven results. Using NCL technology will improve the unbalance of the original data set, and then using a random forest algorithm will make its classification effect better.

6. Conclusion

In summary, this paper lists the results of multiple linear regression, BP neural network, and random forest model. It also provides examples of how commercial property values have changed over time in various places. Each of the three models has pros and cons of its own, and the accuracy of predicting home prices has also grown as a result of resolving each model's issues. To further assure the accuracy of the prediction findings, it is concluded that a range of models can be utilized to forecast the housing price simultaneously.

References

[1] Zhou Liangjin, Zhao Mingyang. Price analysis of second-hand housing in Shenzhen based on Random forest. Chinese market, 2022, (26): 68-71+133.

[2] Song Yao. Research on housing price prediction based on machine learning regression model. Electronic Production,2021, (02): 41-43.

[3] Zhou Liangjin, Zhao Mingyang. Housing Price prediction Analysis based on several types of machine learning models. Circulation of the national economy, 2022, (6): 111-116.

[4] Li Yonghui. Concerning the current effects of house prices and change trend. Dongyue review, 2002, 23(6): 3.

[5] Meng Xianchun. Research on the Correlation Mechanism between Housing Price Fluctuation and Economic Fluctuation in China. Jilin University, 2004.

[6] Jiang Lihong, Li Qinghua. Factors affecting housing location analysis. Journal of urban development, 2005, (4): 3.

[7] Li Ling, Zhu Daolin, et al. Study on the impact of real estate regulation policies on Housing prices based on PSR model -- A case study of Beijing. Resources Science, 2012, 34(4):787-793.

[8] Li Shengda. Housing price forecasting model based on multivariate linear regression. Science and Technology Innovation, 2021, (06): 91-92.

[9] Wang Jing, Luo Weiping, Chen Yongheng. Housing price prediction and analysis based on neural network. Journal of Xiangyang Vocational and Technical College, 2021, (02): 112-115+140.

[10] Li Hanxu, Wei Jiayin, Lu Youjun. Prediction and analysis of second-hand housing price in Shenzhen based on Random Forest. Modern Information Technology, 2021, (15): 100-104.

[11] Luo Bowei, Hong Zhiyong, Wang Jinyi. Application of multiple linear regression statistical model in housing price prediction. Computer Age, 2020, (06): 51-54.

[12] Bu Yanjie. Modern computer of Chengdu Housing price prediction model based on GA-BP neural network, 2022, (20): 98-102.

[13] Wu Qiong, Li Yuntian, Zheng XianWei. Optimization of random forest algorithm for classification of unbalanced training sets. Industrial Control Computers, 2013, (07): 89-90.