

Analysis and Forecast of Housing Demand Impact in China's Underdeveloped Areas Based on XGBOOST Algorithm

Bin Chen^{1,*}, Jia Li²

¹ Arts and Sciences, The University of Alabama, 35401, Tuscaloosa, AL, UA

² Faculty of Journalism, Belarusian State University, 220030, Minsk, Belarusian

*Corresponding author Email: 576812652@qq.com

Abstract. This paper takes China's underdeveloped areas as the research object, and takes Liangshan City, Sichuan as an example, to construct an index system of housing demand and influencing factors. On this basis, the XGBOOST algorithm and test indicators are established. By adjusting the parameters and taking the XGBOOST operation results as an example, combined with the test indicators, the results with the highest accuracy are screened out, and the independent variable importance analysis and housing demand forecast are respectively carried out to provide some inspiration and inspiration for the government. The results show that: (1) The factors affecting housing demand in underdeveloped areas mainly include real estate sales area; population growth and so on. (2) The XGBOOST algorithm constructed in this paper has good robustness and prosperous performance. The R2 is as high as 0.96 and the MSE is only 0.02. It predicts that the future housing demand will show a slow-growth trend.

Keywords: XGBOOST; Predict; Housing demand; Influencing factors.

1. Introduction

Housing system reform is a reform of the housing security system in the process of urbanization based on China's national conditions and is also an important part of China's economic system reform. Looking back on the development process since the reform and opening up over the past 30 years, the housing system policy reform and the changes in the real estate industry closely related to this have produced far-reaching impacts not only in the economic field but also in other important fields. The demand in the residential market directly affects the overall trend of market demand. When studying the demand in the residential market, in addition to fully considering the consumer demand in the residential market, we should also take into account China's current national conditions and policy systems. As a bulk commodity, housing configuration is a dynamic and constantly adjusting process due to the durability of housing. This allows for a constant crossover between allocations and reallocations in the housing market.

This makes it necessary to comprehensively consider the influence of various factors when studying the housing demand. At the current stage in China, the development imbalance between regions still exists. Except for some of the cities that have become international cities and large cities, there are still many underdeveloped areas, and there are still many problems in the demand of the residential market. If the supply is always in the unsaturated stage, the housing demand will not be met, and the market allocation will be chaotic. In the long run, a bubble economy will form, which is not conducive to the healthy development of the real estate market but also has a great negative effect on the national economy. Therefore, based on the xgboost algorithm, it is very necessary to forecast and analyze the impact of housing demand in China's underdeveloped areas. At present, the main methods for forecasting housing demand are: The housing demand forecasting equation $D=h(R) \times N$ proposed by Mankiw and Well (1999), where D is the housing demand, R is the rent, N is the population, $h(R)$ is the average per capita residential demand function, and the change of D is inversely related to the change of $h(R)$; Goh (1999) used multiple regression analysis methods to make regression prediction of residential demand, and the predicted variable is The variables of population size, housing investment scale, national savings, and unemployment rate; David A. Macpherson (1999) used the population analytic hierarchy process to study the problem of housing

demand, and obtained the housing demand through the correlation analysis between variables. The regression equation between the age group of the pre-purchasing population and the prediction ; Bharat Barot (2002) used the error correction in the economic model, the cointegration test, the Granger causality analysis, etc. to calculate the housing demand ; Lux and Sunega (2010) used time series and other indicators to compare and analyze the development trend of housing demand before and after the mortgage crisis in selected countries, establish a model and use the development trend of developed countries after 2007 to revise the model ; Ehsan Shekarian and Alireza Fallahpour (2013) proposed the use of gene expression programming to predict average housing prices, which is also applicable to predicting housing demand ; Kiel and Zabel (2008) used the 3L method to predict housing prices. The demand is divided into three levels, and brought into the Hedonic regression model to analyze the population classification ; S.Thomas Ng and Martin Skitmore (2007) used a genetic algorithm and multiple regression equation to study the housing demand. Question .

Zhao Zhao and Wang Wenyuan (2000) used the total population, the number of households, and the annual number of houses under construction to carry out multiple regression analyses to predict and analyze the development status and development trend of the residential housing market. Zhang Yueping also used multiple regression to predict the demand. Linear relationship with per capita housing area, annual newly registered population, per capita disposable income, urban and rural residents' savings balance, and average housing price ; Geng Jijin (2011) based on Chen Hongyi, Jiang Jingbo (1995) dynamic analysis of urban housing system Model uses system dynamics to divide the real estate system into a land supply subsystem, a housing supply subsystem, a housing price subsystem, and a housing demand subsystem. Song Guoxue (2007) used historical data to apply correlation analysis and other means, and combined Views software to predict the per capita housing area after fifteen years ;

Previous research has discussed housing demand a lot, but in less developed areas and lesser use of machine learning, and this paper provides the need for that. And the previous research is a little more qualitative, and there are some quantitative ones, but they do not involve machine learning. This paper takes China's underdeveloped areas as the research object and takes Xinjiang Province as an example to construct an index system of housing demand and influencing factors. On this basis, the XGBOOST algorithm and test indicators are established. By adjusting the parameters and taking the XGBOOST operation results as an example, combined with the test indicators, the results with the highest accuracy are screened out, and the independent variable importance analysis and housing demand forecast are respectively carried out to provide some inspiration and inspiration for the government.

2. Construction of the indicator system

2.1 Housing demand

In the face of the current situation of excessive real estate growth, excessive land supply, and rapid rise in housing prices, the demand for housing in many underdeveloped areas has changed to a certain extent. The demand for housing also has a great impact on the economic development of these regions and also affects the quality of life of the people in the regions. If the demand for housing is too high, it will lead to a social phenomenon in which the supply exceeds demand, which will cause real estate developers to deliberately raise prices, which is also a major reason for the rise in housing prices, making most people become house slaves. the feeling will also be reduced. If the demand for housing is too low, it will cause a social phenomenon of oversupply. Although it will reduce housing prices, it will also cause a phenomenon of excess housing. Therefore, most contemporary young people choose to rent a house, which will not only reduce living expenses but also increase happiness to a certain extent.

2.2 Index system construction

Economy: The level of economic development of a country or a region is a decisive factor affecting real estate demand. Generally speaking, there is a positive correlation between the level of real estate demand and the development level of the national economy, that is, a country or a region with a high level of economic development has a correspondingly high level of promoting its real estate demand. If the economy is developing fast, then the rate of housing growth in this period will be correspondingly faster. The impact of the economy on the demand for housing is mainly reflected in the following aspects: the level of national income, with the increase of national income with economic development, the expansion and reproduction capacity of enterprises is improved, and the growth of personal disposable income will inevitably increase the demand for real estate. productive and consumer demand. Since China's reform and opening up, the national economy has grown rapidly, which has promoted a sharp increase in the demand for various real estate, thus promoting the prosperity of China's real estate at this stage.

GDP refers to the total market value of new final goods and services produced within an economy over some time. Note that it must be new, so the value of newly built houses will also be included in GDP, and second-hand houses will not be included. Under normal circumstances, cities with high GDP will have higher incomes, and higher incomes will also lead to higher housing prices. The increase in housing prices will eventually lead to changes in housing demand.

Level of urbanization: Urbanization is an inevitable trend of social and economic development. Urbanization includes the increase in the number of cities, the expansion of scale, and the increase in urban population. The increase in the urban population will increase people's demand for housing and will increase the demand for employment arrangements for production and business real estate. At the same time, the development of urban construction requires the renovation of old areas and major construction projects, which will inevitably lead to the demolition and relocation of old areas, resulting in the housing needs of relocated households. To sum up, the index system constructed in this paper is shown in Table 1.

Table 1. System of factors affecting housing demand

First-level indicator	Secondary indicators	Reference
Economy (X1)	GDP(X11)	
	CPI (X12)	
	Exchange rate (X13)	
	Resident per capita income (X14)	
	Price level (X15)	
Real Estate (X2)	House sales area (X21)	
	Real estate price (X22)	
	Per capita living area (X23)	
	Regional Urbanization Process (X31)	
Urban Structure (X3)	Population (X32)	
	Population Growth Rate (X33)	
	Aging situation (X34)	

3. Construction of XGBOOST algorithm

XGBoost (eXtreme Gradient Boosting), also known as the extreme gradient boosting tree, is an implementation of the boosting algorithm. It is an open-source framework for Gradient Boosting created by Dr. Chen Tianqi of the University of Washington. XGB is an implementation of the boosting algorithm, mainly to reduce the bias, that is, to reduce the error of the model. Therefore, it uses multiple base learners, each of which is relatively simple to avoid overfitting. The next learner is to learn the difference between the result of the previous base learner and the actual value. Through the learning of multiple learners, continuously reducing the difference between the model value and the actual value.

$$y_i^2 = f_1(x_i) + f_2(x_i) = y_i^1 + f_2(x_i) \quad (1)$$

$$y_i^t = \sum_{k=1}^t f_k(x_i) = y_i^{t-1} + f_t(x_i) \quad (2)$$

The basic idea is to continuously generate new trees, and each tree learns based on the difference between the previous tree and the target value, thereby reducing the bias of the model. That is, the sum of the results of all trees is the prediction value of the model for a sample. How to choose/generate a better tree at each step is determined by the objective function of this paper.

The objective function consists of two parts, one is the model error, that is, the difference between the true value and the predicted value of the sample, and the other is the structural error of the model, that is, the regular term, which is used to limit the complexity of the model.

$$Obj(\theta) = L(\theta) + \Omega(\theta) = L(y_i, y_i') + k = \sum_{k=1}^t \Omega(f_k(x_i)) \quad (3)$$

Bring it in to the above formula to convert to: $Obj^t = \sum_{n=1}^n L(y_i, y_i^{t-1} + f_t(x_i)) + \Omega(f_t) + \sum_{t=1}^{T-1} \Omega(f_t)$, the

error of the t-th tree consists of three parts, the sum of the errors of the n samples in the t-th tree, the structural error of the t-th tree, and the structure error of the previous t-1 trees. The structural error of the first t-1 trees is constant because this paper already knows the structure of the first t-1 trees. Assuming that the loss function in this paper is the squared loss function (mse), the above objective function is transformed into:

$$Obj^t = \sum_{n=1}^n L(y_i, y_i^{t-1} + f_t(x_i)) + \Omega(f_t) + \sum_{t=1}^{T-1} \Omega(f_t) \quad (4)$$

The above formula is the objective function of the t-the step of GB when the loss function is mse. The only variable is, that the loss function here is still a relatively complex expression, so to simplify it, the second-order Taylor expansion is used to approximate the expression, namely:

$$f(x + \Delta x) = f(x) + f'(x)\Delta x + 1/2 f''(x)\Delta x^2 \quad (5)$$

So let $g_i = \partial_{y_i^{t-1}} l(y_i, y_i^{t-1})$ and $h_i = \partial_{y_i^{t-1}}^2 l(y_i, y_i^{t-1})$ be the first and second derivatives of $l(y_i, y_i^{t-1})$, respectively. Then after the above loss function is converted to the second derivative,

$$Obj^t = \sum_{n=1}^n [l(y_i, y_i^{t-1}) + g_i f_t(x) + 1/2 h_i f_t^2(x)] \quad (6)$$

So when the loss function is MSE $g_i = 2(y_i^{t-1} - y_i)$, $h_i = 2$. After conversion, the first item is the sum of the errors between all samples and the t-1th tree. Because the t-1th tree is known, it can be regarded as a constant item. This article is temporarily in the objective function. By omitting it, the objective function of this paper becomes the function of $f_t(x)$. And C is a function about the output of the leaf node, so the objective functions of this article are all converted to functions about ω :

$$Obj^t = \sum_{n=1}^n [g_i f_t(x) + 1/2 h_i f_t^2(x)] + \Omega(f_t) + constant \quad (7)$$

In this paper, let $G_j = \sum_{i \in I_j} (g_i)$ and let $H_j = \sum_{i \in I_j} (h_i)$ then the objective function of this paper is converted into:

$$Obj^t = \sum_{j=1}^T G_j * \omega_j + 1/2 (H_j + \lambda) * \omega_j^2 + \lambda T \quad (8)$$

In the above expression, j represents the jth node j represents the jth node j represents the jth node, i represents the ith sample. Therefore, the entire objective function is converted into a univariate quadratic function about, that is, the leaf node score. If you want to optimize the objective function, it is to solve the optimal. Therefore, this paper takes the derivation of the objective, and the formula is as follows:

$$\omega^* = -G_j / (H_j + \lambda) \quad (9)$$

Finally, substituting ω^* into the objective function, the objective function becomes:

$$Obj^t = 1/2 \sum_{j=1}^T G_j^2 / (H_j + \lambda) + \lambda T \quad (10)$$

As long as we know the first and second derivatives of the loss function, and which leaf node the sample falls on, this paper only requires that on each leaf node, the first and second derivatives of the sample can be used to calculate the objective function. It can also decide whether to split the node and split according to which node's eigenvalue.

4. Setting of inspection indicators

In the construction of XGBOOST, even if the same software has the same set of data and the same algorithm, many different results are usually obtained due to different parameters selected, and each result is a runnable file. To get a convincing result, this paper introduces test indicators, such as MSE, RMSE, MAE, MAPE, and R^2 . Where MSE (Mean Squared Error): Expected value of the square of the difference between the predicted value and the actual value. The smaller the value, the higher the model accuracy. RMSE (root mean square error): is the square root of MSE, the smaller the value, the higher the model accuracy. MAE (Mean Absolute Error): The average value of the absolute error, which can reflect the actual situation of the predicted value error. The smaller the value, the higher the model accuracy. MAPE (Mean Absolute Percent Error): is the deformation of MAE, it is a percentage value. The smaller the value, the higher the model accuracy. R^2 : Comparing the predicted value to using only the mean, the closer the result is to 1, the more accurate the model is. In this paper, two of the indicators MSE and R^2 can be selected as test indicators for judgment.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (11)$$

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (12)$$

Where y_i is the true value, $\hat{y}_i$ is the predicted value, and $\bar{y}$ is the mean value.

5. Empirical Analysis

5.1 Selection of research objects and data collection

As a relatively poor county in Sichuan Province, Liangshan, Sichuan, currently has a relatively slow population growth and slow economic development. Therefore, taking Liangshan, Sichuan as an example for analysis can be a better case for this study. By collecting the data on the factors affecting housing demand from 2000 to 2020 from the National Bureau of Statistics of China, Juhui Database, and CEIC Database, and after screening and screening, the index system that dominates the impact of housing demand is finally determined. extraction. Using the established index system, a Sichuan housing demand evaluation model based on a random forest is constructed, the parameter optimization model is adjusted, and the prediction set data is substituted into the model to obtain the single-sample relative error. The descriptive statistics of the indicators are shown in Table 2.

Table 2. Indicator Descriptive Statistics

	<i>Mean</i>	<i>Median</i>	<i>Min</i>	<i>Max</i>	<i>Sd</i>
X11	84946.67	78507.00	26014.00	163889.00	42436.78
X12	885966.96	725251.80	134610.30	2186795.89	665333.25
X13	33760.33	29248.00	9220.00	79434.00	19757.06
X14	0.23	0.22	0.28	0.37	0.05
X15	105.00	102.90	94.20	106.60	1.68
X21	466660.62	437593.85	54969.62	1134751.54	343551.28
X22	23459.54	19954.00	8493.49	47358.00	13458.71
X23	64.11	84.95	67.52	77.55	6.33
X31	1850.09	1952.00	1557.00	2589.31	358.90
X32	2585.21	2501.07	552.07	4577.05	1513.89
X33	125.90	53.83	15.01	354.43	125.63
X34	10968.58	10648.50	4355.00	13418.64	2688.36

5.2 Analysis of Influencing Factors

By collecting the data on the factors affecting housing demand from 2000 to 2020 from the National Bureau of Statistics of China, Juhui Database, and CEIC Database, and after screening and screening, the index system that dominates the impact of housing demand is finally determined. extraction. Using the established index system, a Sichuan housing demand evaluation model based on a random forest is constructed, the parameter optimization model is adjusted, and the prediction set data is substituted into the model to obtain the single-sample relative error. Finally, a comparative study was performed using the goodness of fit (R^2) and mean squared error (RMSE). By adjusting the number of leaf nodes, seek the optimal XGBOOST algorithm result. The results show that XGBOOST has the advantages of small prediction error and high model stability. It is a method worthy of application and promotion in real estate appraisal. The data in this paper are shown in Table 3.

Table 3. Prediction results

Leaf Node	R^2	MSE
100	0.129	3.601
200	0.045	0.249
300	0.267	0.214
500	0.642	5.122
1000	0.954	0.136
2000	0.562	0.753
5000	0.759	0.673
10000	0.686	0.168
20000	0.986	0.02

It can be seen that when the number of leaf nodes reaches 20,000, the model accuracy converges, R^2 reaches 0.986, and MSE is only 0.02, indicating that the model has better accuracy and stronger model robustness.

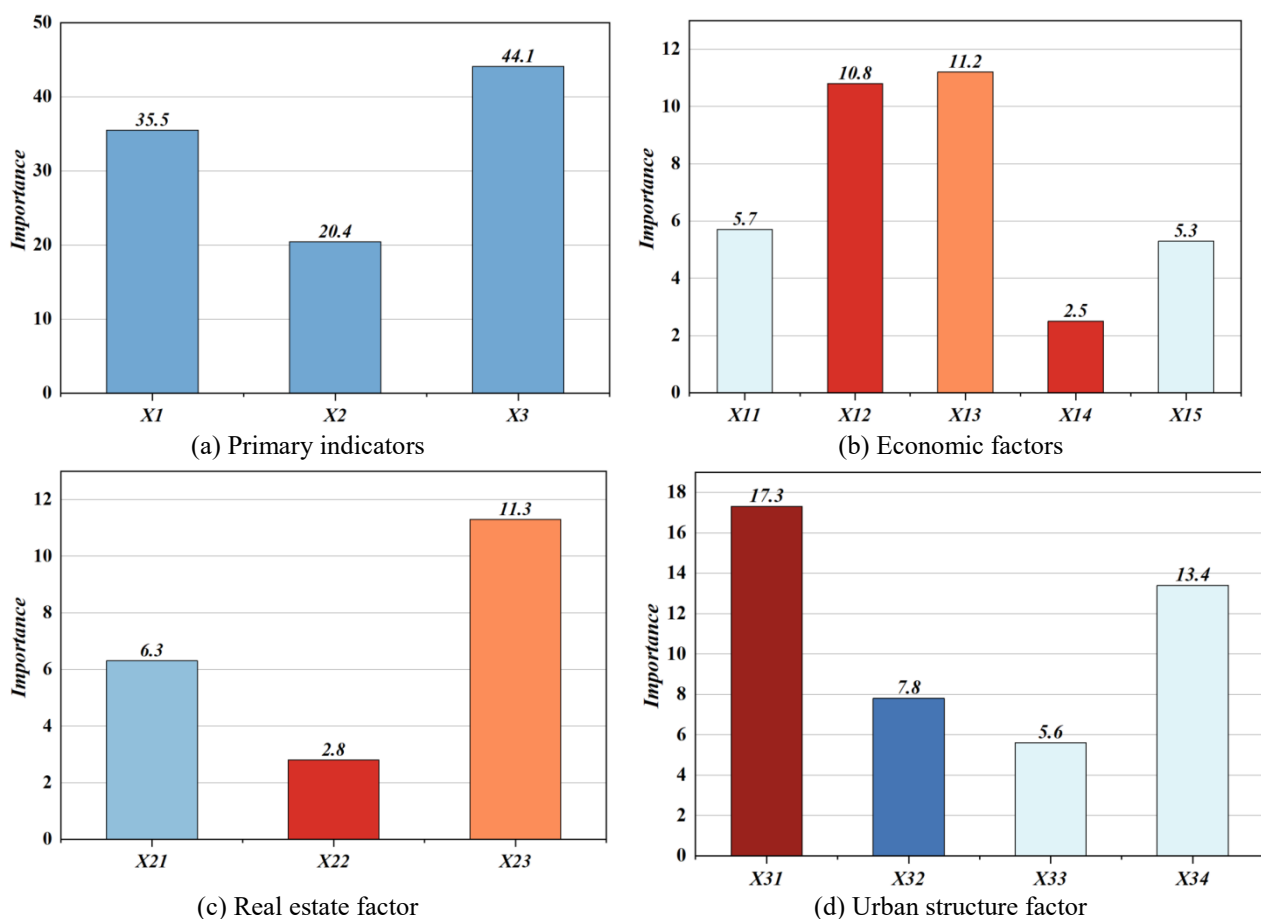


Figure 1. Characteristic size of indicators

As can be seen from Figure 1, in the case of primary indicators, the main factors affecting housing demand are X3, followed by X1, and finally X2. In terms of economy, mainly X12 and X13. In the actual housing demand environment, they are X23, X21, and X22. In terms of urban development, the influencing factors are X31, X34, X32, and X33. Among them, X3 has the largest influencing factor, reaching 44.1, indicating that the actual development of the city determines the influencing factors of housing demand. In addition, X1, reaching 35.5, indicates that the dynamic development of economic factors has seriously affected the logistics in underdeveloped areas. Finally, X2, only 20.4, indicating that the response of urban infrastructure to housing prices is lacking.

5.3 Housing Demand Forecast

Based on the XGBOOST model constructed above, the real estate transaction demand from 2010 to 2021 is selected for forecasting, and the forecast results for the next five years are shown in Figure 2.

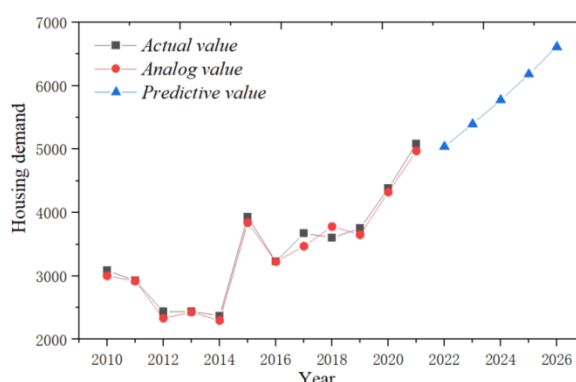


Figure 2. Forecast of housing demand in Liangshan, Sichuan

Based on the SPSSpro platform, the housing demand in Liangshan, Sichuan is predicted in the next five years. The results show that the sales volume will gradually increase in the next five years, reaching 6066 sets of commercial housing in 2026. The increase reached nearly 32 or so. Comparing the main factors of real estate demand in first- and second-tier cities and third- and fourth-tier cities, it is concluded that the regional economic structure is an important factor affecting the regional differences in real estate destocking. To achieve the goal of destocking real estate in first- and second-tier cities, what needs to be solved is to reduce real estate development and the purchasing power of buyers; what needs to be solved in third- and fourth-tier cities to achieve the goal of destocking real estate is the problem of expanding the demand for commercial housing, and improving purchasing power can be achieved by increasing wages. and policy benefits to solve, and increase the demand for commercial housing must be solved by developing the local economy.

The first is to deeply implement regionally differentiated destocking policies. From empirical analysis, it is concluded that the main factors affecting real estate demand in first- and second-tier cities are the average selling price of houses and GDP, and the main factors affecting real estate demand in third- and fourth-tier cities are the balance of RMB urban and rural savings deposits. , the average selling price of houses and the disposable income of all residents. Due to the limitation of land value in first- and second-tier cities, inventory should not be removed. On the other hand, there is more space in third- and fourth-tier cities. Therefore, third- and fourth-tier cities should further implement relevant real estate preferential policies. , and adjust the supply to speed up the process of real estate destocking, and can selectively introduce the household registration system to attract talents. The second is to promote the development of the real economy and ensure high-quality economic development. At present, my country's real estate investment accounts for about 50% of economic growth, which is not conducive to the high-quality development of the economy. It can be seen from the Sino-US trade war that my country's high-tech compared with the United States The industry is relatively lacking, and my country should further accelerate the development of high-tech industries and at the same time promote the development of the real economy. Third, banking

financial institutions should provide a good financial environment for destocking and preventing real estate bubbles. The People's Bank of China should maintain a neutral, stable, and moderately flexible monetary policy. Banking financial institutions should consider the issue of financial risk prevention and control, and implement different implementations according to local conditions. For example, the housing loan policy in first- and second-tier cities should be tightened, the housing loan policy in third- and fourth-tier cities should be moderately flexible and loose, and monetary policy tools should be effectively used to increase destocking support. The fourth is to speed up the transformation of real estate enterprises and build a real estate operation model of "government + bank + enterprise" cooperation, and each local government department should refer to differentiated real estate destocking policies according to their conditions, develop and design rental housing information platforms, and issue relevant documents to regulate rental The housing market, protect the legitimate rights and interests of the renters, ensure the concept of "houses are used for living, not for speculation", and encourage relevant enterprises to invest in the construction of the housing rental market, and support the transformation and upgrading of real estate enterprises.

6. Conclusion

This paper takes the underdeveloped areas of China as the research object, and takes Liangshan City, Sichuan as an example, to construct an index system of housing demand and influencing factors. On this basis, the XGBOOST algorithm and test indicators are established. Through parameter adjustment, taking the XGBOOST operation results as an example, combined with the test indicators, the results with the highest accuracy were selected, and the independent variable importance analysis and housing demand forecast were carried out respectively to provide some enlightenment and inspiration for the government. In summary, the conclusions are as follows:

(1) Factors affecting housing demand in underdeveloped areas mainly include real estate sales area; population growth and so on.

(2) The XGBOOST algorithm constructed in this paper has good robustness and good performance. R^2 is as high as 0.96 and MSE is only 0.02. It is expected that the housing demand will show a slow growth trend in the future. And in the 2026 year, the demand for housing in Liangshan County, Sichuan will reach 6,600 units.

References

- [1] Wu Qiang, Tu Kun, Zeng Yifan, etc. Build my country's main energy (coal) later discussion on the main problems and countermeasures faced by the advanced version [J]. Chinese Journal of Coal, 2019, 44(6): 1625-1636.
- [2] Mao Dongxin. The influence of coal-based primary energy consumption structure on my country[D]. Xiamen: Xiamen University, 2019.
- [3] Chen Xiangguo. The development of the coal industry should focus on policy orientation [J]. Festival Energy and Environmental Protection, 2019, 000(2): 42-43.
- [4] Lu Ming. Research and application of China's coal consumption forecasting model [D]. Beijing: North Beijing Jiaotong University, 2018.
- [5] Zhang Jinsuo, Feng Xue, Zou Shaohui. my country's coal demand forecast based on-trend combination Research on measurement model [J]. Business Research, 2014(6): 51-56.
- [6] Wang Lijie, Sun Jihu. Coal demand forecast model based on grey system theory [J]. Chinese Journal of Coal, 2022(3): 333-336.
- [7] Xie Heping, Wu Lixin, Zheng Dezhi. China's energy consumption and coal demand in 2025 Prediction [J]. Chinese Journal of Coal, 2019, 44(7): 1949-1960.
- [8] Yan Wang, Jingwen L. China's Present Situation of Coal Consumption and Future Coal Demand Forecast [J]. China Population Resources and Environment, 2018, 18(3): 152-155.

- [9] Zhu Q, Zhang Z, Li R, et al. Structural Analysis and Total Coal Demand Forecast in China [J]. *Discrete Dynamics in Nature and Society*, 2014: 1-10.
- [10] Bates J M, Granger C W J. The Combination of Forecasts [J]. *Journal of the Operational Research Society*, 2019, 20(4):451-468.
- [11] Zhang Youlan, Zhou Aimin. Optimal weighted combination prediction and its application [J]. *Quantity Economic Technology and Economic Research*, 2017(10): 66-68.
- [12] Geng Kui. Discussion on the optimal weighted combination method [J]. *Quantitative Economics Technical Economics Research Research*, 2018, 15(11): 44-46.
- [13] Yang Yingming, Li Quansheng, Fang Jie. Coal consumption based on optimal weighted combination model Predictive analysis of fees [J]. *Coal Engineering*, 2018, 50(5): 156-160.
- [14] Yan Xiaohong. Prediction model of energy consumption combination based on IOWGA operator [J].
- [15] *Journal of Southwest University for Nationalities: Natural Science Edition*, 2019(4): 53-57.
- [16] Lu Zhanhai. Research on coal consumption forecasting based on a combined forecasting model [J]. *China Coal*, 2018, 38(11): 12-15.
- [17] Zhang Enying, Zhang Xiaohan. Application of ARIMA Model in Heilongjiang GDP Forecast [J]. *Shopping mall modernization*, 2014(3): 187-188.