Proceedings of the 3rd International Conference on Applied Physics and Mathematical Modeling
DOI: 10.54254/2753-8818/2025.DL27031

Applying Hypothesis Testing and XGBoost to Analyze and Predict Housing Price Trends in Guangzhou

Junyang Liu

School of Mathematics (Zhuhai), Sun Yat-sen University, Zhuhai, China
3040144308@qq.com

Abstract. With the increasing application of machine learning technologies in the real estate sector, traditional models face challenges in terms of prediction accuracy and capturing nonlinear relationships. This study aims to integrate classical statistical methods with advanced machine learning algorithms to thoroughly explore the core factors influencing housing prices in Guangzhou and achieve high-precision regional price predictions. By employing statistical hypothesis testing to preliminarily screen factors, the significance of their relationships with housing prices is verified. On this basis, an XGBoost learning model is constructed to perform modeling and prediction, leveraging its strong capabilities in handling nonlinear relationships, interaction effects, and missing data. The study identifies the number of new houses, GDP per capita, the number of subway stations, and the number of large supermarkets as key factors verified to have a significant positive impact on housing prices in Guangzhou. Housing prices in Guangzhou's core districts have taken the lead in recovery, while development zones have shown steady growth with the greatest potential. In contrast, the peripheral suburban areas continue to face downward pressure.

Keywords: Housing Price Prediction, Univariate regression equation, Hypothesis Testing, XGBOOST Model

1. Introduction

In recent years, with the booming real estate market and persistently high housing prices, real estate valuation has become a topic of widespread concern in social life. Accurately predicting future housing price trends is of great significance for economic development and people's living standards. As a first-tier city in China and a core engine of the Guangdong-Hong Kong-Macao Greater Bay Area, Guangzhou has always attracted significant attention regarding its housing price movements. A decade ago, driven by the nationwide property market boom, housing prices across various districts in Guangzhou also experienced considerable increases. Over the years, Guangzhou's housing market has exhibited clear regional differentiation and structural characteristics, and predictive research has continued to evolve in response to market changes. The Guangzhou real estate market is characterized by its high complexity and dynamic nature, with prices determined by a combination of numerous linear and nonlinear factors. Traditional linear models and logistic regression exhibit significant limitations in capturing these intricate relationships. As a representative gradient boosting decision tree algorithm, XGBoost has gained

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2025.DL27031

widespread popularity due to its outstanding performance in both predictive competitions and practical applications. This study employs hypothesis testing based on linear regression and the XGBoost model to conduct a comprehensive analytical process—from "feature screening" to "accurate prediction." It investigates the factors influencing housing prices in Guangzhou and, after excluding less impactful variables, forecasts future price trends. The research aims to provide the public with insights into which factors drive changes in Guangzhou's housing market and to offer predictions regarding its future direction, thereby presenting a data-driven solution for housing price forecasting in the city.

2. Literature review

Zhao proposed six factors that affect house prices, which are date of transaction, age of the house, distance to the nearest subway station, number of convenience stores around the house, latitude, and longitude [1].

Wang suggested several factors to predict house price, including interest rate at the beginning of the year, interest rate at the end of the year, growth rate of construction area, growth rate of completed area, growth rate of sales area, index of gross domestic product per capita, growth rate of consumption per capita, and growth rate of house price [2]. And Zheng and his team confirmed the significant value-added effect of subway stations on house prices using rigorous econometric methods [3]. Wang and his team found that two core factors, GDP per capita and population size, have a significant impact on house prices in large and medium-sized Chinese cities at the macro level [4].

Long in his article empirically investigated the impact of the distribution and accessibility of various types of urban public service facilities on residential house prices by taking Beijing as an example, which included factors affecting house prices such as educational facilities, commercial service facilities, medical facilities, transportation facilities, and public service facilities [5].

3. Methodology

3.1. Introduction to the experiment

3.1.1. Factor screening and significance test

In this paper, the six independent variables of consumption growth rate, the number of new houses, GDP per capita, the number of subway stations, the number of hypermarkets, and the birth rate of the population and the dependent variable of regional house prices are respectively subjected to significance tests. Clarify which factors are statistically significant influencing factors, and preliminarily determine the direction of their influence (such as the number of new homes may be negatively correlated with house prices).

3.1.2. XGBoost prediction model construction and performance evaluation

Five districts of Guangzhou, Tianhe, Yuexiu, Haizhu, Zengcheng, and Huadu, are taken as sample units, and their quantitative data corresponding to the six factors and current house price data are collected. The data were divided into training and testing sets to construct the XGBoost prediction model, and the root mean square error (RMSE) and other indicators were used to evaluate the model's prediction accuracy of the house prices in each district in Guangzhou to ensure the effectiveness of the model.

DOI: 10.54254/2753-8818/2025.DL27031

3.2. Specific methods

3.2.1. Univariate regression equation

A regression model is used to analyze how the dependent variable y depends on the independent variables x_1 , x_2 , x_3 ,… x_k and the error termε. Its general form is as follows [6]:

$$y_i \ \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ \ldots \ + \beta_k x_{ki} + \varepsilon_i \qquad \ldots \tag{1}$$

where b_1 b_2 b_3 are the parameters of model, and ε_i is the error term, which reflects the effect on y of random variables other than the linear relationship between x_1 , x_2 , x_3 ,… x_k and y .

When parameter estimation is performed by the least squares method, which is based on the principle of solving the parameter estimates by minimizing the sum of squares of the residuals [6]:

$$\min \ \sum_{i=1}^{n} \varepsilon_i^{\ 2} \ \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \ldots - \beta_k x_{ki})^2 \tag{2}$$

Deriving Q with respect to β_0,β_1, β_2 …… $\beta_k$respectively yields the following equation:

$$\frac{\partial Q}{\partial \beta_0} \ 2 \sum (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \ldots - \beta_k x_{ki}) (-1) = 0 \tag{3}$$

$$\frac{\partial Q}{\partial \beta_1} = 2 \sum (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \ldots - \beta_k x_{ki}) (-x_{1i}) = 0 \tag{4}$$

$$\frac{\partial Q}{\partial \beta_k} = 2 \sum (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \ldots - \beta_k x_{ki}) (-x_{ki}) = 0 \tag{5}$$

The parameter estimates of β_0, β_1, β_2 … β_k are the solutions of the above equation.

3.2.2. Hypothesis testing

3.2.2.1. T-test

In addition to the test of the regression model as a whole, this paper also needs to evaluate the significance of each β_i , which can be adjusted according to the performance of each indicator in terms of the number of indicators and the form of indicators. t-test is mainly used to observe whether the explanatory variable x_j is linearly related to the dependent variable y. The original hypothesis $H_0: \beta_i = 0$. The formulas that may be used include: Covariance matrix of coefficients $\mathrm{Cov}(b) = \hat{\sigma}^2 (X^T X)^{-1}$, standard error $SE(b_j) = \sqrt{\mathrm{Cov}(b)_{jj}}$, t-statistic $t_j = \frac{b_j}{SE(b_j)}$, and p−value $p_j = 2 \times P(T_{n-p} > |t_j|)$ [7].

DOI: 10.54254/2753-8818/2025.DL27031

3.2.2.2. F-test

The F-test is commonly used to test the overall significance of a regression model. If the original hypothesis $H_0\ \beta_1=\beta_2=\beta_3=\ldots=\beta_k=0$ is true, then it indicates that there is no linear relationship between y and $x_1, x_2, x_3, \ldots x_k$ as a whole, and all the variation in y comes from the disturbance term ε, and the value of the F-statistic is small. If the original hypothesis $H_0\ \beta_1=\beta_2=\beta_3=\ldots=\beta_k=0$ is false, then it indicates that y is linearly related to $x_1, x_2, x_3, \ldots x_k$ as a whole, and that $x_1, x_2, x_3, \ldots x_k$ has a significant effect on y, resulting in a larger F statistic. Significance F represents the critical value of F, i.e., the P value. It is used to represent the probability that the model is false, then true is represented by 1-P. So the smaller the P-value the better. Some of the equations that may be used are sum of squares $SST=\sum(y_i-\bar{y})^2$, regression sum of squares $SSR=\sum(\hat{y}_i-\bar{y})^2$, F-statistic $F=\frac{SSR/p}{SSE/(n-p-1)}$ and p-value $P(F_{p,n-p-1}>F)$ [7].

3.2.3. XGBoost model

XGBoost (eXtreme Gradient Boosting) is an optimized distributed gradient boosting library, which is based on gradient boosting decision trees, introduces regularization terms and second-order derivative information to improve the performance, and supports parallel computation, sparse data processing, and multiple distributed environments. The system excels in classification, regression and other tasks through technical innovations such as weighted histogram algorithm and cache optimization, and is widely used in data competitions and industrial fields.

The core of XGBoost is Objective Function Optimization, which is divided into two parts: loss function (prediction error) and regularization (control model complexity).

The expression of the objective function is

$$Obj=\sum_{i=1}^{n} L(y_i,\hat{y}_i)+\sum_{k=1}^{K}\Omega(f_k) \tag{6}$$

Where $L(y_i,\hat{y}_i)$ represents the loss of the i th sample (e.g., mean square error, cross entropy), $\Omega(f_k)$ represents the complexity penalty of the kth tree (to prevent overfitting). $\hat{y}_i$ represents the prediction results of the first K trees cumulatively.

The expression for the complexity (regularization term) of the tree is

$$\Omega(f)=\gamma T+\frac{1}{2}\lambda\|\omega\|^2 \tag{7}$$

Where T is the number of leaf nodes of the tree(the fewer the simpler),w is the weight of the leaf nodes (prediction value), andγ andλ are hyperparameters that control the penalty strength.

The expression of XGBoost approximating the objective function with a second-order Taylor expansion (which is more accurate than the first-order of traditional GBDT) is

$$Obj\approx\sum_{i=1}^{n}[L(y_i,\hat{y}_i)+g_i f_t(x_i)+\frac{1}{2}h_i f_t^2(x_i)]+\Omega(f_t) \tag{8}$$

DOI: 10.54254/2753-8818/2025.DL27031

Where $g_i=\partial_{\hat{y}_i^{t-1}} L\left(y_i,\hat{y}_i^{t-1}\right)$ represents the first-order gradient and $h_i=\partial^2_{\hat{y}_i^{t-1}} L\left(y_i,\hat{y}_i^{t-1}\right)$ represents the second-order gradient.

The optimal weight for the j th leaf of the t th tree can be obtained by derivation as

$$w_j^*=-\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i+\lambda} \tag{9}$$

where I_j represents the set of samples belonging to the j th leaf [6].

3.3. Experimental process

This study collected data from five districts of Tianhe, Yuexiu, Haizhu, Zengcheng and Huadu in the past 11 years, and realized the one-dimensional regression analysis and hypothesis test through MATLAB to judge which factor has a significant influence on house price, screen out irrelevant factors and then use MATLAB to make the future house price prediction with the homemade XGBOOST model. The following are some of the data:

Table 1 intercepts the "Guangzhou Real Estate Market Analysis Report" released by the Guangzhou Housing and Urban-Rural Development Bureau (GZHURD) in the past years or the relevant market data from the Guangzhou Real Estate Industry Association (GZREIA), which summarizes the number of new houses in the five districts of Guangzhou.

Table 1. Number of new housing districts

Time	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
Tianhe	25	28	32	35	30	28	25	22	28	26	24	25
Yuexiu	5	4	6	5	4	3	3	2	3	2	2	2
Haizhu	18	20	22	25	22	20	18	15	20	18	20	22
Zengcheng	42	48	55	65	70	75	78	85	80	75	70	65
Huadu	35	40	45	52	58	60	62	65	60	58	55	52

As shown in Table 2, this paper intercepts the data of "Sub-district GDP" and "Sub-district Resident Population" in the Guangzhou Statistical Yearbook of past years, and uses the formula of GDP per capita to calculate the per capita GDP of each sub-district in Guangzhou from 2013 to 2024. GDP per capita.

DOI: 10.54254/2753-8818/2025.DL27031

Table 2. GDP per capita

Time	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
Tianhe	270,009	291,024	315,323	341,475	370,620	301,757	322969	334287	370498	386125	401500	420000
Yuexiu	193,340	207,918	223,565	240,392	258,800	288,536	308,567	318,005	335,102	348,506	362,000	378,000
Haizhu	114,007	122,379	131,418	141,247	151,914	163,181	175,420	181,205	201,588	210,658	220,000	230,000
Zengcheng	68,152	73,605	79,493	85,852	92,720	99,108	106,958	110,805	118,965	124,258	130,000	136,000
Huadu	96,882	104,632	113,002	122,042	131,805	142,350	153,737	159,120	169,003	176,562	184,000	192,000

According to the official opening announcement of Guangzhou Metro Group, the number of metro stations in the five districts of Guangzhou were collected and organized from 2013 to 2024 (Table 3).

Table 3. Number of metro stations

Time	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
Tianhe	22	22	24	24	26	28	28	28	30	30	31	33
Yiuxiu	19	19	19	19	19	19	19	19	19	19	19	19
Haizhu	15	15	18	18	21	23	23	23	26	26	28	29
Zengcheng	0	0	0	0	0	0	4	4	4	4	4	5
Huadu	2	2	2	2	2	2	2	2	2	2	2	2

Table 4 comprehensively analyzed the “Report on the Development of Retail Industry in Guangzhou”, commercial real estate databases such as Winsome Business, and the official website of the stores of major supermarkets, and compiled the number of large integrated supermarkets with an area of over 6000 square meters in the five districts of Guangzhou during the past ten years, which are located in the five districts. 6,000 square meters or more of large supermarkets (such as Wal-Mart, Carrefour, China Resources Vanguard, etc.).

Table 4. Number of Large Supermarkets

Time	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
Tianhe	18	19	20	21	22	23	23	23	22	22	21	21
Yuexiu	15	15	16	16	16	16	16	16	15	15	15	15
Haizhu	12	13	13	14	14	15	15	15	15	15	16	16
Zengcheng	4	5	5	6	7	8	9	10	11	12	13	14
Huadu	5	6	6	7	8	9	10	10	11	11	12	12

As shown in Table 5, this paper collates the average house prices in the five districts of Guangzhou for each year through the “Operation of the Real Estate Market in Guangzhou” report released by the Guangzhou Municipal Bureau of Housing and Urban-Rural Development on a monthly basis.

DOI: 10.54254/2753-8818/2025.DL27031

Table 5. Average house prices

Time	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
Tianhe	38000	40500	42000	47800	53000	58000	65000	72000	85000	88000	90000	95000
Yuexiu	35500	37000	38500	44000	48500	52000	56000	60000	70000	72000	75000	78000
Haizhu	30000	32000	33500	40000	46000	52000	58000	65000	75000	78000	82000	85000
Zengcheng	8500	9000	9200	11500	14500	17500	19500	20500	21000	18500	17000	16500
Huadu	9000	9200	9500	11000	13800	16200	18000	19000	19500	17000	16000	15800

3.4. Results

Table 6 demonstrates the values of the coefficients of b1 to b6 solved by MATLAB and the results of the t-test

Table 6. One-way regression coefficients and t-test results

	Coefficient	StdError	tStat	pValue
b1(consumption growth rate)	-184.01	335.48	-5.4851	0.58583
b2(number of new houses)	449.95	86.671	5.1915	3.999e-06
b3(GDP per capita)	0.16201	0.023926	6.7714	1.4995e-08
b4(number of subway stations)	3264.8	350.04	9.3269	1.9309e-12
b5(number of hypermarkets)	-3454.6	739.85	-4.6692	2.3805e-05
b6(birth rate)	-653.99	394.18	-1.6591	0.10348

After the F-test, the F-statistic is 94.3091, and the P-value of the F-test is infinitely close to zero.

Table 7 shows the predicted house prices in the five districts of Guangzhou after implementing the XGBOOST model with MATLAB,

Table 7. Predicted house prices in five districts in 2024

Tianhe	Yuexiu	Haizhu	Zengcheng	Huadu
89798	75740	74474	28103	15097

The above predicted results reveal that among the six factors of house prices in Guangzhou, the core factors are the number of new houses, GDP per capita, the number of metro stations and the number of large supermarkets, and in the prediction of house prices, there is a huge difference between the core districts and periphery districts, and the house prices of Tianhe District are still far ahead, while the house prices of Huadu and Zengcheng District remain lower. Tianhe district is still far ahead, while Huadu and Zengcheng districts keep their prices low, probably because they are the furthest away from the main city and have a large number of new houses.

4. Discussion

There are significant differences in the influence of the six selected factors affecting Guangzhou house prices. Among them, the number of new houses, GDP per capita, the number of metro stations, and the number of hypermarkets are verified as the key factors that have a significant positive impact on the house prices in Guangzhou, which represent the economic purchasing power

DOI: 10.54254/2753-8818/2025.DL27031

base of the region, the convenience of transportation, and the maturity of the living facilities, which together constitute the three pillars of the core value of the property, and the GDP per capita is ranked the highest in terms of its significance, which further highlights that economic fundamentals are the most fundamental force to support the house prices. The significance of metro and commercial support also reaffirms that the capitalization of public service value is a stabilizing mechanism to shape the spatial pattern of house prices; while the hypothesis test and XGBoost model of this study show that the impact of consumption growth rate and birth rate on the current house price in Guangzhou does not pass the statistical significance test.

Based on the XGBoost model, the predicted prices of the five districts in Guangzhou in the future clearly show the differences in regional prices and the characteristics of market differentiation. The predicted price of Tianhe District reaches RMB 89,839 / ㎡, far ahead of other districts, which confirms its absolute position as the new economic center and financial core of Guangzhou, with high GDP per capita, dense metro network and top commercial facilities supporting its high price level; the predicted price of Yuexiu District is very close to that of Haizhu District, which is RMB 76,449/㎡ and 76,500/㎡ respectively, although there may be differences in the factors driving the price of the two districts. However, there may be differences in the factors driving the prices of the two districts, Yuexiu District, with the advantage of the traditional education district, has a deep historical deposits, while the Haizhu District, thanks to the rise of the Pazhou Internet Cluster and the development opportunities brought about by the urban renewal, showing great development potential; Zengcheng District, the forecast price of 23,800 yuan/㎡, significantly lower than the core area, the price level reflects its status as a large supply of This price level reflects its status quo as a “house price depression” with a large supply and far distance from the core area, in which the large number of new houses is the main factor inhibiting its house price; Huadu District has the lowest projected price of RMB 15,206/㎡, which is in line with its status quo of being the furthest from the main urban area in terms of commuting distance, and with a relatively low GDP per capita, placing it at the bottom end of the value echelon of Guangzhou's real estate market. This is in line with its distance from the main city and relatively low GDP per capita, placing it at the end of the value ladder in Guangzhou.

5. Conclusion

Taking the Guangzhou real estate market as the research object, this study aims to penetrate the fog of market fluctuations, scientifically identify the core driving factors affecting house prices, and realize data-driven forecasts of future trends. Through a combination of statistical hypothesis testing and XGBoost machine learning modeling, the study reveals that GDP per capita, metro accessibility, maturity of commercial support and market supply and demand (number of new homes) are the key dimensions determining the differentiation of house prices in various regions of Guangzhou at the current stage. Especially importantly, the study finds that the consumption growth rate and the birth rate lack significant explanatory power for short-term house prices, a finding that challenges conventional perceptions and reveals the complexity and laggedness of the real estate market.

The empirical analysis shows that Guangzhou's property market is undergoing a profound revaluation, and its future is far from a simple “general increase” or “general decrease”, but rather a dramatic structural differentiation. The predicted prices of the Tianhe, Yuexiu, Haizhu, Zengcheng and Huadu districts clearly outline a picture of core areas with strong values and peripheral areas under pressure. This study determines the degree of influence of several factors affecting house prices, which can help the public rationally consider purchasing a home.

DOI: 10.54254/2753-8818/2025.DL27031

Although this study provides a valuable exploration of the factors influencing house prices in Guangzhou through hypothesis testing and XGBoost modeling, and draws some illuminating conclusions, there are still limitations in that the data in this study are mainly based on macro statistics at the district or board level, and such aggregated data can hide the great variations within regions. In addition, by focusing on the six core factors with high data availability, this study inevitably omits other potential variables that are not included in the model but are equally important. This variable omission problem is a common limitation of feature price models, and indeed most machine learning models, and it leads to a lack of comprehensiveness in the model's explanation of house prices.

Future research can improve the model by refining the granularity of the data and moving the unit of analysis from the “district” down to the “street” or even the “building”. Collect the listing information and transaction data of each neighborhood on the trading platform, so as to improve the accuracy and practicality of the model. In addition, key variables such as policies and market sentiment are included to build a more comprehensive analytical framework, and real estate news and government planning documents are included in the model as proxies to quantify market expectations.

Despite the limitations of this study in terms of methodology, data granularity and scope of analysis, these shortcomings precisely point out a clear path for subsequent studies. The significance of this study lies not only in its conclusions, but also in the analytical framework it provides, which provides the necessary quantitative basis and decision-making reference for policymakers and market participants, with a view to helping Guangzhou's real estate market move towards a more rational, healthy, and stable future.

References

[1] Zhao Tinghan. Analyzing factors affecting house prices based on multiple regression model [J]. China Science and Technology Investment, 2020(04): 129-131.
[2] WANG Hao. Application of Logistics Regression Model in Predicting House Prices in Guangdong Province [J]. Modern Business, 2010(16): 304-306.
[3] Zheng Siqi, Fu Yuming, Ren Rongrong. (2014). Impact of urban rail transit on residential prices and land development intensity - A case study of Beijing [J]. Journal of Geography, 69(2), 111-120.
[4] Wang Yang, Wang Zhenpo, Zhao Yanxia. (2017). Per capita GDP, population size and urban house prices - An empirical study based on panel data of 35 large and medium-sized cities in China [J]. Exploration of Economic Issues, (3), 80-86.
[5] LONG Fenjie, WU Jing, ZHENG Siqi. A study on the impact of urban public service facilities on residential prices--Taking Beijing as an example [J]. Urban Development Research, 2009, 16(11): 76-81.
[6] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785 -794).
[7] J. P. Jia, X. Q. He, Y. J. Kim. Statistics (8th ed.) [M]. Beijing: Renmin University of China Press, 2021.
[8] Song Bo, Sun Weizeng, Zheng Siqi. Urban public services and housing prices: Estimation based on boundary breakpoint regression [J]. Economics (Quarterly), 2022, 22(1): 193-212.