International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 9 Issue 5, Sep-Oct 2025 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Machine Learning-Based Beijing Housing Price Prediction System

Zihao Wang, Hao Zhang, Tianshuang Han, Xiuyi Yang, Yinuo Liu

Student, School of Information, Beijing Wuzi University, Beijing, China

ABSTRACT

With rapid socioeconomic development, the real estate market faces dual challenges of increasing homebuying pressure and rising investment risks. As a unique commodity combining residential attributes and asset value, the formation mechanism of real estate prices has grown increasingly complex. It is not only influenced by multiple factors but also exhibits significant nonlinear characteristics in its overall trend. As a first-tier city in China, Beijing's housing prices are influenced by various factors, including economic policies, geographical location, educational resources, and transportation accessibility. Accurately forecasting housing price changes holds significant importance for homebuyers, investors, real estate developers, and government policymakers. This project aims to leverage machine learning and data analytics to propose a multi-source data fusion framework. By collecting local environmental data such as air quality and noise pollution, it breaks the boundaries of traditional real estate datasets. For the first time, dynamic environmental indicators are incorporated into the housing price evaluation system. This approach systematically quantifies the impact mechanism of human living environment quality on real estate value. A baseline model, XGBoost, is constructed to handle high-dimensional features, establishing a model capable of predicting Beijing housing prices. This assists users in understanding price trends and supports decision-making.

How to cite this paper: Zihao Wang | Hao Zhang | Tianshuang Han | Xiuyi Yang | Yinuo Liu "Machine Learning-Based Beijing Housing Price Prediction System" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-9 | Issue-5, October 2025, pp.389-392, URL: www.ijtsrd.com/papers/ijtsrd97504.pdf

IJTSRD97504

KEYWORDS: *Machine Learning; Housing Price Prediction; XGBoost.*

INTRODUCTION

China's real estate market has experienced rapid growth over the past few decades, becoming a vital pillar of the national economy. As the capital, Beijing serves not only as a political and cultural center but also as a core hub for economy, technology, and finance, making its housing price fluctuations highly representative. Since the 1998 housing market reforms, Beijing's housing prices have generally trended upward, yet significant market volatility has emerged due to policy interventions, economic cycles, and population mobility. These fluctuations indicate that Beijing's housing prices are influenced by multifaceted factors—including policy, economy, society, and demographics-making prediction challenging yet academically valuable.

Real estate price forecasting employs diverse methodologies, with most approaches relying on mathematical statistical models grounded in economic principles, predominantly linear regression models. However, housing price trends exhibit complex and variable characteristics, typically displaying pronounced nonlinearity while being influenced by multiple intersecting factors. This complexity leads to significant discrepancies between linear model predictions and actual market fluctuations.

Regarding the predictive analysis of housing prices and their influencing factors, Wang Huiwen[1] employed a multiple linear regression model to establish the regression relationship between the two. Gu Xiujuan et al.[2] treated the dynamic evolution of housing prices as a Markov chain, forecasting Beijing's housing price trends to provide consumption guidance for homebuyers while offering reference suggestions for government regulatory policy formulation. Yang Nan et al.[3] employed a grey Markov model to conduct empirical forecasting research on Shanghai's residential and office building indices. For Taiyuan housing price data, Hou Puguang[4] utilized wavelet analysis theory for decomposition and reconstruction, establishing an ARIMA model through parameter estimation to

achieve housing price forecasting. Yu Baolu et al.[5] constructed an improved BP neural network model by optimizing input nodes, initial weight selection, and activation functions. Li Daying et al.[6] first screened key factors influencing real estate prices using rough set theory, then performed predictive analysis on real estate price indices based on wavelet neural networks. Yao Yao et al.[7] integrated convolutional neural networks (CNN) with random forest (RF) algorithms. By combining housing price data with remote sensing imagery, they constructed a mining model that achieved precise micro-level mapping of housing prices without additional data support. Li Chunsheng et al.[8] Leveraging the global search capability of genetic algorithms to optimize the initial weights and thresholds of BP neural networks, two housing price prediction models were established: a traditional BP neural network and an improved GA-BP neural network. Li Shengda[9] A multivariate regression analysis model for housing price prediction was proposed, incorporating unemployment rate, loan interest rate, and national consumption index as variables, demonstrating its superior data fitting performance. Ding Fei et al.[10] By integrating the migration mechanism and spiral search mechanism of the lion pack algorithm with the seagull algorithm, combined with indicators such as housing layout and area, a predictive study on second-hand housing prices in Qingdao City was conducted. Chen Zekun et al.[11] established a regression analysis and prediction model based on the gradient descent algorithm using Boston housing price data as a sample. Jiang Zhongyun et al.[12] integrated the Keras framework's ELU and linear activation functions with the RMSprop optimization method into a BP network, constructing a Shanghai second-hand housing price prediction model based on Lianjia housing price data. In international research, Osland et al.[13] improved upon models like geogewichtete Regression and introduced spatial econometric models. Rondineri et al.[14]estimated rental price trends using rental price data provided by real estate developers and census data.

This study constructs a deep learning-based housing price prediction system. It selects air quality indicators, noise environment quality parameters, and the number of infrastructure facilities such as bus/subway stations, medical institutions, educational institutions, and parks/green spaces across different regions of Beijing as model input features. The model undergoes iterative training using the XGBoost algorithm to fit the true distribution characteristics of the data, ultimately achieving housing price prediction functionality. The system supports user interactions such as property price prediction queries and related topic discussions, providing valuable reference for homebuyers' decision-making processes.

Research Design

This study primarily employs quantitative analysis methods for in-depth processing and analysis of collected data. First, descriptive statistical analysis calculates metrics such as mean, median, standard deviation, and frequency to summarize central tendency, dispersion, and distribution characteristics, providing an intuitive and comprehensive overview of Beijing's housing market fundamentals. For instance, when analyzing the impact of economic growth on Beijing's housing prices from 2015 to 2019, the average economic growth rate during this period was approximately 6.68% with a standard deviation of about 0.302%. Meanwhile, the average housing price growth rate was approximately 11.96% with a standard deviation of about 23.37%. This indicates a positive correlation between economic growth and Beijing's housing prices. Second, machine learning and multivariate statistical methods-such as linear regression and LSTM-were employed to test research hypotheses, explore relationships between variables, and assess the significance of differences. By constructing linear regression models, the extent of independent variables' influence on dependent variables and their predictive capabilities were analyzed. Finally, feature variables were selected for housing price forecasting. The specific experimental research methods and steps are outlined below:

1. **Data Acquisition:** This study established a "web crawler-API-database" integrated data acquisition system, enabling systematic collection of multidimensional data sources through technological convergence. A distributed crawler cluster was built using the Scrapy 2.6.2 framework within the Python ecosystem, combined with a Redis database for distributed scheduling of URL task queues. The data parsing layer utilizes BeautifulSoup 4.11.1 for HTML document tree analysis, complemented by the lxml parser (C-based kernel) to enhance DOM node locator efficiency. For JavaScript-dynamic rendered pages (e.g., Lianjia.com's housing price trend charts), Selenium 4.4.3 integrated with ChromeDriver was employed to simulate headless browser loading.

Data Sources and Crawling Strategies:

Housing Transaction Data: Targeted crawling of Anjuke's national historical housing price column (2024-2025, monthly frequency), locating target data blocks via XPath expressions.

Environmental Quality Data: Retrieve daily AQI indices, PM2.5 concentrations (μg/m³), and noise

decibel levels (daytime/nighttime) for each district from the Beijing Municipal Ecology and Environment Bureau's "Real-Time Air Quality Release Platform." Utilize direct JSON interface connections to bypass page rendering delays.

Infrastructure POI Data: Utilizing the POI search interface from Amap's open platform web services, facilities such as subway stations, educational institutions, and medical facilities are retrieved via keyword searches. Returned results include multiple attributes including name, address, latitude/longitude, and POI type.

2. **Model Construction:** This study employs Extreme Gradient Boosting Tree (XGBoost) as the baseline model for housing price prediction. Leveraging its strengths in high-dimensional feature handling, nonlinear relationship fitting, and computational efficiency, a customized model architecture achieves comprehensive modeling of factors influencing Beijing housing prices.

XGBoost (eXtreme Gradient Boosting) is a gradient boosting decision tree ensemble framework proposed by Chen & Guestrin (2016). Its core innovation lies in balancing model performance and generalization capability through a regularized objective function and an efficient node splitting algorithm. The model's objective function is defined as: $L(\phi)=i=1\sum nl(yi,y\hat{}i(t-1)+ft(xi))+\Omega(ft)$, where $l(\cdot)$ denotes the loss function (mean squared error MSE is adopted in this study), $\Omega(ft)=\gamma T+21\lambda j=1\sum Twj2$ serves as the regularization term (where n is the number of leaf nodes in the tree and wj is the weight of the leaf node). By approximating the loss function via a second-order Taylor expansion, the model can be transformed into a quadratic optimization problem for each leaf node weight, significantly enhancing convergence speed.

Conclusion

This project successfully developed a Beijing housing price prediction and analysis system based on Python and machine learning technologies, achieving the three core objectives:

1. Prediction Accuracy: The model achieved an accuracy rate of 87%, exceeding the 85% target, and effectively predicted housing prices across districts.
2. Attribution Analysis: Quantitatively identified "school district resources" as the most critical factor influencing Beijing housing prices, followed by geographic location, subway accessibility, and property attributes.
3. Visual Presentation: Prediction outcomes, price distributions, and ranking of influencing factors were clearly displayed through charts.

Limitations:

1. Data Limitations: Insufficient timeliness of data and granularity of key information like school districts; lacks subjective factors such as orientation and renovation quality.
2. Model Limitations: Difficulty in predicting sudden policy changes; potential reduction in generalization capability for distant suburban areas.
3. System Integration: Currently functions more as an analytical prototype than a mature, interactive application system.

Future Directions:

1. Establish real-time data pipelines to incorporate multidimensional data.
2. Explore advanced models (e.g., LSTM for time-series forecasting) and develop region-specific models.
3. Develop interactive web applications and enhance model prediction interpretability.

6. References

[1] Wang Huiwen, Meng Jie. Predictive Modeling Methods for Multiple Linear Regression [J]. Journal of Beijing University of Aeronautics and Astronautics, 2007, (04): 500-504.

[2] Gu, X. J., & Li, C. (2012). Research on housing price prediction based on Markov chains [J]. Consumer Economics, 28(5), 40-42.

[3] Yang Nan, Xing Lichong. Application of Grey Markov Models in Predicting Housing Price Indices [J]. Forum of Statistics and Information, 2006, 21(5): 52-55.

[4] Hou Puguang, Qiao Zequn. Research on Housing Price Forecasting Based on Wavelet Analysis and ARMA Models [J]. Statistics and Decision Making, 2014, 17(15): 20-22.

[5] Yu Baolu, Duan Xun, Wu Yun. Establishment and Application of BP Neural Network Data Forecasting Model [J]. Computer and Digital Engineering, 2016, 44(3): 482-486.

[6] Li Daying, Xu Wei, Chen Rongqiu. Research on Real Estate Price Trend Forecasting Based on Rough Set and Wavelet Neural Network Models [J]. Management Review, 2009, 21(11): 18-22.

[7] Yao Yao, Ren Shuliang, Wang Junyi, et al. Micro-scale Mapping Method for Urban Housing Prices Using Convolutional Neural Networks and Random Forests [J]. Journal of Geoinformation Science, 2019, 21(2): 168-177.

[8] Li Chunsheng, Li Xiaoye, Zhang Kejia. BP Neural Network-Based Housing Price Forecasting Analysis Enhanced by Genetic Algorithms [J]. Computer Technology and Development, 2018, 28(8): 144-147.

[9] Li Shengda. Housing Price Prediction Model Based on Multivariate Linear Regression [J]. 2021, 91(2): 91-92.

[10] Ding Fei, Jiang Mingyan. Housing Price Prediction Based on an Improved Lion Pack Algorithm and BP Neural Network Model [J]. Journal of Shandong University (Engineering Edition), 2021, 51(3): 1-9.

[11] Chen Zekun, Cheng Xiaorong. Regression Analysis and Prediction of Housing Prices Based on Gradient Descent Algorithm [J]. Information Technology and Informatization, 2020(5): 10-13.

[12] JIANG Zhongyun, SHEN Guoxin, Prediction of House Price Based on The Back Propagation Neural Network in The Keras Deep Learning Framework [C]//ICSAI, 2019: 1408-1412.

[13] Osland L. An Application of Spatial Econometrics in Relation to Hedonic House Price Modeling [J]. Journal of Real Estate Research, 2010, 32(3): 289-320.

[14] Rondinelli C, Veronese G. Housing Rent Dynamics in Italy [J]. Questioni Di Economia E Finanza, 2010, 28 (62): 540-548.