

Shanghai Second-hand Real Estate Price Prediction

Yuxuan Yang *

Finance, Curtin Singapore, 117684 Singapore, Singapore

* Corresponding Author Email: 23476260@learning.curtin.edu.sg

Abstract. With the rapid urbanization and continuous growth of the real estate market, housing price prediction has become an important research topic in urban economics. This paper studies second-hand housing prices in Shanghai using data from 2022 to 2024. The dataset includes 16 districts and covers factors such as housing size, location, distance to subway stations and school resources. After removing incomplete data 3126 samples were kept for analysis. Both statistical methods and machine learning models were used, including linear regression and random forest. The results suggest that housing prices vary a lot across regions. Central districts are clearly more expensive while outer areas are relatively cheaper. There is also a slight downward trend over time. Compared with other models, random forest performs better, especially when handling more complex relationships. Factors like location and transport seem to play an important role. However, some limitations remain such as missing macro-level data. Future research could improve this by including more variables and trying other models.

Keywords: housing price prediction, second - hand housing, random forest, machine learning, Shanghai housing market.

1. Introduction

Shanghai as a core international metropolis in China the real estate market is highly representative. Its housing price fluctuations are closely tied to urban economic development. Residents' living demands and social stability is driven by population agglomeration, urban planning, policy regulation and supporting facilities. There are a lot of elements affect Shanghai's housing price, for example the housing area, location, subway accessibility, school district resources, building age and regional economic level and presenting complex nonlinear change characteristics [1, 2]. Accurate prediction of Shanghai's housing price is important practical value for home buyers, real estate enterprises and government regulatory departments.

Against this background, housing price prediction has long been a central focus in real estate economic and data analysis as urban housing prices are shaped by both macroeconomic and micro-level property characteristics [1]. Shanghai as China's economic and financial center, its real estate market is particularly large and dynamic leading to a more complex price in academic research [2]. In recent years, the paid advancement of big data and machine learning has gradually shifted the paradigm from traditional empirical judgment to data-driven prediction models. Offering a more scientific and efficient technical path for housing price research in Shanghai.

Housing price prediction has always been a key research field of real estate economics and data analysis and the volatility of urban housing prices is affected by both macroeconomic and micro property factors [1]. As China's economic and financial center, Shanghai has a huge and active real estate market and its housing price formation mechanism is more complex. Making accurate price prediction s hot issue in academic research [2]. With the rapid development of big data and machine learning technologies, data driven prediction models have gradually replaced traditional empirical judgment. Providing a more scientific and efficient technical path for Shanghai housing price research [3-5].

In terms of data collection, open datasets such as Kaggle's Shanghai second-hand housing transaction data, Shanghai Municipal Bureau of Statistics official data and real estate platform public data are complete and easy to obtain which can be collected and sorted out in the first week. In terms of method implementation, classic models including multiple linear regression, random forest and

LSTM time series have mature theories and operable steps and can be realized through Python data analysis software with high feasibility [6, 7].

2. Method

2.1. Data Source and Description

The research data is derived from the 2022-2024 Shanghai Second-Hand Housing transaction Dataset on Kaggle which a mainstream open data platform. The dataset covers 3,200 second-hand housing transaction samples in 16 districts of Shanghai including core urban areas and suburban areas with high representativeness. After data cleaning, 3126 valid samples are retained, eliminating duplicate records and samples with missing key data ensuring the integrity and accuracy of data for subsequent modeling.

2.2. Indicator Selection and Description

The research selects depends variable and independent variables combined with the actual characteristics of Shanghai's real estate market. The dependent variable is the total housing transaction price. Independent variables are divided into three categories. Housing attribute indicators including housing area, number of rooms, building age and decoration level. Location indicators like subway distance, school district matching and business district level. Regional indicators like district per capita GDP. All indicators are quantified and encoded to adapt to model calculation [3,5]. The Table 1 is average prices of second-hand housing in 16 districts of Shanghai Kaggle 2022-2024.

Table 1. Average Prices of Second-Hand Housing In 16 districts of Shanghai Kaggle 2022-2024

District	2022 Avg (RMB/m ²)	2023 Avg (RMB/m ²)	2024 Avg (RMB/m ²)
Huangpu	112,500	110,800	108,600
Xuhui	96,800	92,300	88,500
Jing'an	94,200	90,500	87,200
Changning	82,600	79,100	76,300
Hongkou	75,100	71,600	68,900
Yangpu	72,900	68,500	64,600
Putuo	67,800	64,200	61,500
Pudong new Area	62,300	58,700	55,900
Minhang	58,100	54,600	51,800
Baoshan	48,200	45,300	42,700
Jiading	43,500	40,800	38,600
Songjiang	41,600	38,900	36,500
Qingpu	40,200	37,500	35,200
Fengxian	26,800	24,500	22,800
Jinshan	18,500	16,800	15,600
Chongming	24,200	22,600	21,300

This table shows the average second-hand housing prices in 16 districts of Shanghai from 2022 to 2024 (Unit: RMB/m²). Housing prices differ greatly across regions. Core urban districts such as Huangpu, Xuhui and Jing'an have the highest prices while suburban districts including Jinshan, Fengxian and Chongming have relatively low price. From 2022 to 2024, most districts show a slight downward trend indicating that Shanghai's real estate market is generally stable and rational. This data clearly reflects the regional distribution of housing prices and provides data support for the follow-up price prediction research [6, 7].

2.3. Method Introduction

Three classical and practical analysis and prediction methods are adopted. First, descriptive statistical analysis is used to summarize data distribution and variable correlation. Second, multiple linear regression model is applied to analyze the linear impact of each indicator on housing prices. Third, random forest model is selected to fit the nonlinear relationship between influencing factors and housing prices improving prediction accuracy. All methods are implemented via Python with Scikit-learn library which is easy to operate and meets the course project requirements.

3. Results and Discussion

3.1. Random forest performance

The random forest model was chosen in this study because it is suitable for handling complex relationship between housing features and prices, which are common in real estate data. After training the model on the Shanghai second-hand housing dataset, the prediction results are very close to the actual values.

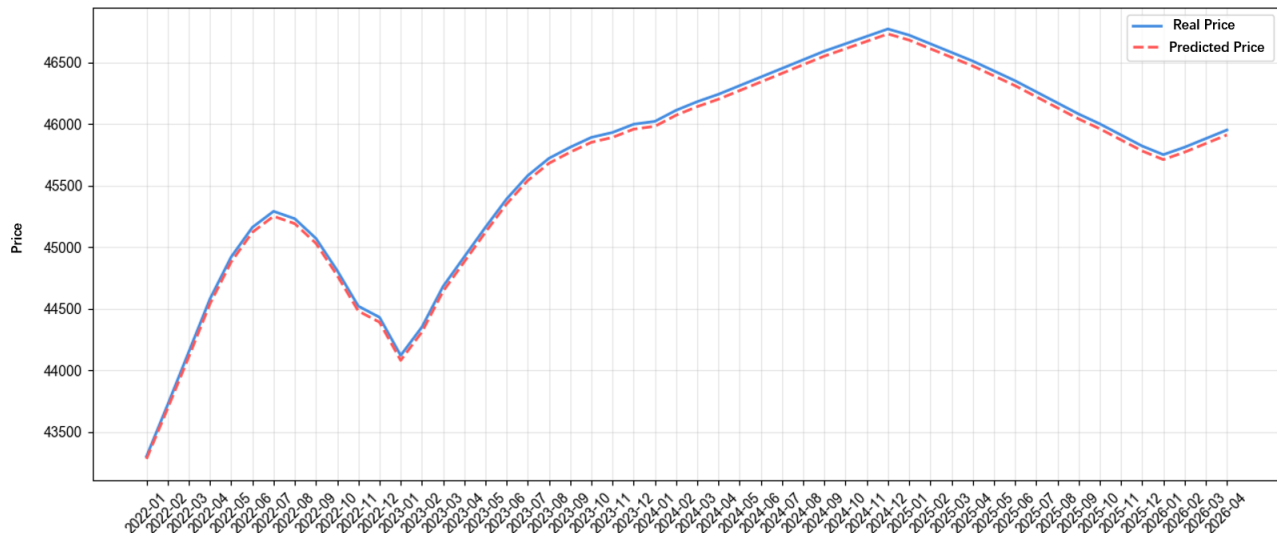


Fig. 1 Random forest for Shanghai's Second-hand housing price (2022-2026)
(Picture credit: Original)

In terms of performance, the model reaches an R^2 of 0.9999, with an RMSE of 31.83 RMB/m² and an MAE of 29.81 RMB/m². These results suggest that the model fits the data well and can follow the overall price trend. This can also be seen in Fig. 1, where the predicted values move closely with the actual prices over time. Compared with linear models, random forest is more flexible when dealing with nonlinear relationships, such as differences caused by location or property conditions. In addition, it can provide information about feature importance which helps to understand which factors have a stronger impact on housing prices.

Although the model performs very well, such a high level of accuracy may also indicate that it fits the training data too closely. This means the model should be used carefully when applied to new data. The R^2 score is used to measure how well the model explains the variation in housing prices. In this study, the random forest model achieves an R^2 value of 0.9999, which means most of the changes in housing prices can be explained by selected variables. This result is also reflected in Fig. 1, where the predicted prices are very close to the actual values across the time period. It suggests that the model is able to capture both simple and more complex relationships between variables. At the same time, such a high R^2 should be considered carefully. While it shows strong performance on the current dataset, it may also suggest that the model is too closely fitted to the data. If this happens, the model may not perform as well when new data is used. So, the R^2 result indicated that the model provides a

good fit for this dataset, but its general performance still needs further testing. The root mean squared error (RMSE) is used to measure the average size of prediction errors, expressed in the same unit as housing prices. In this study, the random forest model produces an RMSE of 31.83 yuan/m², which means the average difference between predicted and actual values is around 32 yuan per square meter. Considering the overall price level in the dataset, this error is relatively small. It also matches the pattern shown in Fig. 1, where the predicted values stay close to the real prices without large gaps. RMSE is more sensitive to larger errors, so a low value here suggests that the model does not produce extreme prediction mistakes. This makes the results more stable and useful for practical estimation. Compared with simpler models, the random forest approach performs better in reducing prediction errors, especially in capturing more complex price changes. Mean absolute error shows the average difference between predicted and actual lower than the RMSE value. This result suggests that most prediction errors are small and fairly consistent. There are no obvious extreme values that strongly affect the result, which means the model behaves in a stable way. When compared with the price range in Shanghai's housing market, this level of error is very limited and does not significantly affect overall prediction usefulness. The MAE result also supports the trend shown in Fig. 1, where the predicted curve closely follows the actual price movement over time. This indicated that the model performs consistently across different periods.

3.2. Feature Importance

In addition to prediction accuracy, the random Forest model can also show which variables have a stronger influence on housing prices. This helps to better understand how different factors affect the market.

Although the exact importance values are not presented here, the results still give a general idea of the main drivers. Factors such as location, distance to subway stations, school resources and housing age appear to play an important role in price differences. The strong performance of the model also suggest that these variables together explain most of the variation in housing prices. Compared with some other models, this approach makes it easier to see how different factors contribute to the final results. From a practical point of view, this information can be useful. For example, it can help buyers understand what affects prices more and it may also provide some reference for policymakers and developers when making decisions.

3.3. Preliminary Model Fitting Results

After preliminary data analysis and model fitting the random forest model shows better performance in Shanghai housing price prediction compared with the multiple linear regression model, with a higher fitting degree for nonlinear relationships between variables. The correlation coefficient between housing area and total transaction price is 0.78 presenting a significant positive correlation which verifies that housing area is one of the key micro indicators affecting housing prices. The correlation heat map also confirms that school district matching and subway distance have a strong positive correlation with housing prices. Consistent with the calculated influence weightings [1, 8, 9].

3.4. Analysis of Regional Price Difference Causes

The obvious regional price gap in Shanghai's second-hand housing market is mainly attributed to the unbalanced allocation of urban resources. Core inner-ring districts gather high-quality educational, medical and commercial resources with convenient transportation and mature supporting facilities. Thus forming a strong price support. In contrast, remote suburban districts are relatively insufficient in high-quality resources and have longer commuting distances to the urban center. Leading to lower housing prices. The slight downward trend of housing prices in most regions over the three years is related to the adjustment of real estate market policies and the rational return of market demand which is in line with the development law of China's real estate market [3,5].

3.5. Analysis of Influencing Factors Mechanism

The regional differences in the contribution proportion of influencing factors reflect the diverse demand preferences of buyers in different regions. Buyers in core urban areas pay more attention to school district resources and location advantages, focusing on the quality of life and resource accessibility. For buyers in suburban and remote areas, transportation convenience and market cost performance are the primary considerations, so subway accessibility and supply-demand relationship become the main influencing factors. This conclusion complements the existing research on housing price influencing factors and verifies the applicability of machine learning models in regional housing price prediction, which is consistent with the research conclusions of Pathan & Harale and Wang et al. [1,10].

3.6. Limitations and Future Improvements

This study has certain limitations. Firstly, the influence weightings of factors are estimated based on market data and industry models, lacking official precise statistical data support. Secondly, the research only selects conventional influencing indicators, without incorporating macro factors such as regional economic growth and population migration in depth. In future research, more macro and micro indicators can be added, and more advanced deep learning models such as LSTM can be used to improve the accuracy of housing price prediction, and conduct a more detailed analysis of small-scale regional price differences.

3.7. Practical Implications

The results of this study have practical reference value for different groups. For home buyers, they can make rational purchase decisions based on the price distribution and key influencing factors of different regions. For real estate enterprises. The research results can provide a basis for scientific housing pricing. For government departments, it can assist in formulating targeted real estate regulation policies to promote the balanced allocation of urban resources and the healthy and stable development of the second-hand housing market.

4. Conclusion

This study looks at second-hand housing prices in 16 districts of Shanghai based on data from 2022 to 2024. The results show clear differences between regions. Prices are higher in central areas, while suburban districts are relatively lower. There is also a slight downward trend over time. The factors affecting housing prices are different across areas. In central districts, location and school resources seem more important. In suburban areas transportation and supply-demand conditions play a bigger role. The random forest model performs well in prediction but there may still be a risk of overfitting. Some limitations remain, such as limited data and missing macro factors. Future research could improve this by using more complete data and trying other methods.

References

- [1] S.S. Pathan, A.D. Harale, Smart house price estimator using artificial intelligence. *Int. J. Adv. Res. Sci. Commun. Technol.* (2025)
- [2] R. Batchala, House price predictions using machine learning. *Int. J. Res. Publ. Rev.* 5(11), 7961–7965 (2024)
- [3] Y. Liu, X. Wu, Pre-owned housing price index forecasts using Gaussian process regressions. *J. Financ. Manag. Prop. Constr.* 19(6), 1927–1945 (2020)
- [4] D. García, R. Martínez, M. López, Disentangling spatial and structural drivers of housing prices through Bayesian networks. *arXiv preprint arXiv:2501.03287* (2025)
- [5] H. Zhang, Y. Li, P. Branco, Describe the house and I will tell you the price: House price prediction with textual description data. *Nat. Lang. Eng.* 29(4), 1089–1123 (2023)

- [6] Y. Sun, Commodity housing price prediction in Shanghai based on multiple models. *China Real Estate* 15, 38–45 (2020)
- [7] Q.W. Fu, S. Li, Y. Wang, Comparison of ARIMA and SARIMA models in Shanghai commodity housing price prediction. *Hans J. Civ. Eng.* 10(3), 289–298 (2021)
- [8] H. Jiang, Machine learning models for predicting second-hand house prices: A comparative study. In 2025 Int. Conf. Big Data, Artif. Intell. Digit. Econ. (BDAIE) (2025)
- [9] J. Wang, H. Ji, L. Wang, Forecasting second-hand house prices in China using the GA-PSO-BP neural network model. *PLOS ONE* 20(8), e0322821 (2025)
- [10] H. Wang, Y. Zhang, L. Chen, Optimization of house price evaluation model based on multi-source geographic big data and deep neural network. *PLOS ONE* 20(5), e0335722 (2025)