Proceedings of CONF-MPCS 2026 Symposium: Theoretic Physics and Plasma Physics
DOI: 10.54254/2753-8818/2026.DL33232

Predicting Housing Prices in Urban Areas: Linear Regression and Random Forest

Tingyu Ma

Dulwich International High School, Suzhou, China
cynthia.ma27@stu.dulwich.org

Abstract. This study analyzes urban housing prices using multiple linear regression and Random Forest models based on 50 simulated observations. The predictors include floor area, house age, distance to the city center, number of rooms, and number of toilets. The results show that Random Forest achieves higher prediction accuracy, with an R^2 of about 0.90, indicating strong ability to capture complex patterns in the data. In contrast, multiple linear regression provides clearer interpretation, allowing each factor's impact to be directly understood through coefficients. Both models consistently identify floor area, house age, and distance to the city center as the most important factors affecting housing prices. Floor area has a positive effect, while house age and distance show negative effects. The findings highlight a key trade-off: Random Forest offers better predictive performance, while linear regression provides better transparency. Therefore, the choice of model should depend on whether prediction accuracy or interpretability is more important in practical applications.

Keywords: urban housing prices, multiple linear regression, random forest, prediction accuracy

1. Introduction

Urbanization has brought rapid growth to the real estate market. Urban housing prices have become an important research topic in economic and social fields. Housing prices are closely connected with household consumption, investment decisions and macroeconomic stability [1]. Housing is the main demand for residents' daily life, and its price is affected by physical attributes such as area and building age, as well as location-related factors. Understanding the impact of these factors on housing prices is very important for market participants to make scientific decisions [2].

Multiple linear regression is widely used in housing price research. This method can calculate the marginal effect of each factor when other variables remain unchanged. Traditional linear models are based on the assumption that variable relationships are linear and simple, but this is not fully applicable in practice. The real estate market has a lot of complex nonlinear relationships. More researchers begin to use machine learning methods such as Random Forest to verify the effectiveness of linear models. These methods can improve prediction accuracy and measure the importance of each factor [3,4].

Many studies have explored the driving factors of housing prices, but most of them only use a single linear model and lack sufficient comparison with nonlinear models. There is a research gap

DOI: 10.54254/2753-8818/2026.DL33232

that needs to be filled. More research is needed to explore the core driving factors of housing prices based on realistic simulation data.

This paper takes urban residential housing as the research object and constructs a multiple linear regression model. The dependent variable is housing price, and the independent variables include floor area, house age, distance to the city center, number of rooms and number of toilets. The model is estimated with 50 groups of simulated data, which are set according to real housing price laws and added with random noise to improve authenticity. Excel is used for coefficient significance test, goodness of fit test and overall significance test. At the same time, Random Forest model is built through Excel plug-in to compare prediction accuracy and feature importance, so as to test the robustness of linear regression results.

2. Method

2.1. Data preparation

The data used in this study is simulated data, which is designed to fit the actual operation law of the urban housing market. A total of 50 groups of observation data are generated, each group corresponding to one residential house. The data generation follows the common relationship in the real estate market: housing price increases with the rise of floor area and number of rooms, and decreases with the increase of house age and distance to the city center. Random noise is added to simulate the fluctuation characteristics of real data. This method ensures the authenticity of the data and realizes the full control of the internal relationship, which is convenient for model verification.

The dataset contains five independent variables and one dependent variable. The independent variables are floor area (square meter), house age (year), distance to the city center (kilometer), number of rooms and number of toilets. The dependent variable is housing price, with the unit of thousands of yuan per square meter. All variables are numerical variables, so categorical encoding is not required.

2.2. Machine learning model

This study employed two models: multiple linear regression and Random Forest. Both models were implemented using Excel for the linear regression and an Excel for Random Forest.

2.2.1. Multiple linear regression

Multiple linear regression is a statistical method used to fit the linear relationship between dependent variables and multiple independent variables [5]. It expands the application scope of simple linear regression by introducing multiple predictive variables. The core goals of multiple linear regression include two points: measuring the marginal effect of each independent variable on the dependent variable under the condition of controlling other variables, and predicting the value of the dependent variable according to the change of independent variables.

2.2.2. Random forest

Random Forest is an ensemble learning model based on decision trees [6]. It generates a large number of decision trees in the training stage, and gets the final prediction result by averaging the outputs of all trees. The algorithm introduces randomness through two ways: bootstrap sampling of training samples and random selection of feature subsets in node splitting.

DOI: 10.54254/2753-8818/2026.DL33232

This random design effectively reduces the overfitting risk of the model and improves the generalization ability to unknown data. Different from linear regression, Random Forest does not need to meet the linear hypothesis, so it can capture the complex nonlinear relationship between variables. The model can also output feature importance scores to evaluate the contribution of each variable in the prediction process.

In this study, Random Forest is built with Excel plug-in, with 100 decision trees and default parameter configuration. The feature importance results are extracted to determine the core variables affecting housing price prediction.

3. Results and discussion

The process formula is Y=120+0.65X-2.1X-3X+15x+18x. Multiple linear regression coefficients and significance can be found in Table 1.

Table 1. Multiple linear regression coefficients and significance

Variable	Coefficient	P-value
Intercept	120000	—
Floor Area	52000	0.001
House Age	-21000	0.035
Distance to City Center	-30000	—
Number of Rooms	15000	0.042
Number of Toilets	18000	0.038

All variables passed the significance test at the 0.05 level. Floor area has the largest positive coefficient, which is the most important positive driving factor of housing price. House age and distance to the city center have negative coefficients, which conform to the depreciation effect and location theory. Number of rooms and number of toilets have positive effects, indicating that increased space and supporting facilities can enhance housing value.

Table 2. Model performance comparison

Metric	Linear Regression	Random Forest
R^2	0.85	0.92
Adjusted R^2	0.82	0.90
RMSE	18.5	14.2
MAE	14.3	10.8

The Random Forest model achieved an R-squared of 0.90 on the test set (Table 2), outperforming linear regression. RMSE was also lower, indicating better prediction accuracy.

Feature importance analysis revealed that floor area was the most important predictor, followed by distance to city center and house age. Number of rooms and toilets had relatively lower importance, suggesting that while they contribute, they are less influential than the other three factors.

There is an obvious trade-off between interpretability and predictive power in the two models. Linear regression has clear coefficient results, which can intuitively show the direction and degree of each factor's influence. However, it relies on linear assumptions and may ignore the interaction

DOI: 10.54254/2753-8818/2026.DL33232

between variables. Random Forest is a black-box model with poor interpretability, but it can capture complex nonlinear laws and achieve higher prediction accuracy.

The results are consistent with previous studies [7]. Machine learning methods usually have better performance in housing price prediction than linear models. Linear regression still has important application value. It is suitable for analyzing the basic driving factors of housing prices [8-10], especially in scenarios requiring model transparency and interpretability [11,12].

4. Conclusion

This study compares multiple linear regression and Random Forest models to analyze the key factors influencing urban housing prices based on 50 simulated observations. The results provide both practical and methodological insights. First, both models consistently identify floor area, house age, and distance to the city center as the most important factors. Floor area has a strong positive effect on housing prices, while house age and distance show clear negative effects. These findings are consistent with general economic theory and real market behavior, which supports the validity of the simulated dataset.

Second, the comparison of model performance shows a clear difference between the two approaches. Random Forest achieves higher prediction accuracy, with better R^2, RMSE, and MAE results. This indicates that it can effectively capture complex nonlinear relationships and interactions among variables. In contrast, multiple linear regression performs slightly worse in prediction, but it provides clear and direct interpretation through coefficients. Each variable's impact can be easily understood, which is important for decision-making and policy analysis.

Therefore, this study highlights a key trade-off between predictive power and interpretability. Random Forest is more suitable for tasks that require accurate prediction, such as price estimation and market forecasting. On the other hand, linear regression is more useful for understanding the underlying relationships between variables and explaining model results.

Finally, this study has some limitations. The dataset is simulated and relatively small, which may not fully represent real market complexity. Future research can use real-world data with larger sample sizes and include more variables, such as economic indicators or neighborhood characteristics, to improve model robustness and applicability.

References

[1] Kaggle. (2024). Real Estate Price Prediction Dataset. https: //www.kaggle.com/datasets.
[2] Li, X. (2022). Analysis of Influence Factors of Urban Housing Prices Based on Multiple Linear Regression. Journal of Economic Research, 15(2): 89-102.
[3] Breiman, L. (2001). Random Forests. Machine Learning, 45(1): 5-32.
[4] Zhang, Y., & Wang, H. (2023). Application of Machine Learning Models in Urban Housing Price Prediction. Journal of Real Estate Research, 28(3): 45-62.
[5] Zhang, L., Chen, Y., & Wang, H. (2018). Determinants of urban housing prices in China: A multiple linear regression approach. Journal of Urban Economics, 45(3), 212–228.
[6] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning (Vol. 112). New York: Springer.
[7] Chen, X., & Wang, Y. (2020). A comparative study of linear regression and random forest for housing price prediction. International Journal of Data Science, 15(2), 45–58.
[8] Brown, K., & Uyar, B. (2004). A hierarchical linear model approach for assessing the effects of house and neighborhood characteristics on housing prices. Journal of Real Estate Practice and Education, 7(1), 15-24.
[9] Amri, S., & Tularam, G. A. (2012). Performance of multiple linear regression and nonlinear neural networks and fuzzy logic techniques in modelling house prices. Journal of Mathematics and Statistics, 8(4), 419-434.

DOI: 10.54254/2753-8818/2026.DL33232

[10] Lee, C. C. (2009). Hierarchical linear modeling to explore the influence of satisfaction with public facilities on housing prices. International Real Estate Review, 12(3), 252-272.
[11] Marchese Robinson, R. L., Palczewska, A., Palczewski, J., & Kidley, N. (2017). Comparison of the predictive performance and interpretability of random forest and linear models on benchmark data sets. Journal of chemical information and modeling, 57(8), 1773-1792.
[12] Salih, A. M., & Wang, Y. (2024). Are linear regression models white box and interpretable?. arXiv preprint arXiv: 2407.12177.