A machine learning driven approach on evaluating second-hand house price, heterogeneity test, and optimization path

Aihua Han[a], Zirui Hu[a], Jiaying Liu[a], Qitong Zhao[b], Fan Zhang[c,d,*]

[a]*School of Statistics and Mathematics, Zhongnan University of Economics and Law, 182 Nanhu Avenue Wuhan, 430073, Hubei Province, China*
[b]*Columbia University, 116th St New York, 10027 New York, United States*
[c]*Ocean College, Zhejiang University, 1 Zheda Road Zhoushan, 316000, Zhejiang, China*
[d]*Kavli Institute for Astrophysics and Space Research Center, Massachusetts Institute of Technology, 185 Albany Street Cambridge, 02139, MA, U.S.A*

Abstract

This study examined factors influencing housing prices using a dataset of 2799 second-hand units. Eleven key factors were analyzed with five machine learning methods including regression trees and random forests. The random forest model identified influential variables and detected heterogeneity across samples. Regression trees and partial dependence plots verified pricing optimization pathways. Primary determinants of price were found to be location, floor area, construction area, and proximity to schools. Urban area, floor level, and household type had significantly heterogeneous price impacts. Regression trees confirmed factor importance while partial dependence plots detailed impacts of floors, construction areas, nearby schools, and bathrooms. It is recommended governments strengthen market oversight and implement differentiated price controls considering local conditions. Buyers should account for variable importance to avoid overpaying.

Keywords: Second-hand house price, Random forest algorithm, Model comparison, Bias effect graph

[*] Fan Zhang is the corresponding author.

[**] Aihua Han, Zirui Hu, Jiaying Liu and Qitong zhao contributed equally to this work and should be considered co-first authors

Email address: f.zhang@zju.edu.cn (Fan Zhang)

1. Introduction

Rapid economic growth and diminishing developable land have increased activity in secondary housing markets. In major cities like Shanghai and Beijing, existing home transactions exceed 70% of total sales. Small-medium cities also see existing homes dominating. Wenzhou housing prices peaked during a 2008 real estate boom but burst, plunging the city into a downturn with new purchase policies until stabilizing in 2015. Transaction volumes for existing Wenzhou urban homes have risen gradually since 2011. By 2020, 31,281 existing home sales rose 3.6% totalling 30.61% of transactions. Accurately predicting existing home prices is vital for stable secondary market development, necessitating predictive model comparisons and factor analysis influencing prices.

With rapid real estate growth, batch evaluation of residential prices is increasingly important domestically and internationally. Regarding existing home pricing models, some studies apply statistical estimations while others achieve good results by leveraging data mining and machine learning techniques. Beyond model selection, some incorporate characteristic variables influencing transactions and big data era insights for more comprehensive and accurate price evaluations.

In terms of researching model selection, (Antipov and Pokryshevskaya, 2012) first applied random forests to housing price evaluation, comparing it to models like regression, trees, and neural networks, finding random forests performed best. (Foryś, 2022) believed not enough attention was paid to machine learning for housing price evaluation, assisting multivariate regression and machine learning to automatically and accurately estimate many property attributes quickly. (Lahmiri et al., 2023) explored machine learning models for housing price forecasting, using enhanced regression trees, support vector regression, and Gaussian process regression. Bayesian optimization determined optimal kernels and parameters, finding improved regression trees performed best, followed by Gaussian processes and then support vector regression.

(Park and Bae, 2015) developed a housing price prediction model using machine learning algorithms like C4.5, RIPPER, Naive Bayes, and AdaBoost, comparing classification accuracy. Results showed RIPPER performed better than other forecasting models. (Ho et al., 2021) evaluated real estate prices using support vector machines, random forests, and gradient-boosted machines. The study found machine learning promising for valuation and appraisal. Additional work based on random forests and recurrent neural networks, (Afonso et al., 2019) show that enriching data sets and combining different machine learning methods may be a better choice for predicting house prices. (Adetunji et al., 2022) explored random forests for house price forecasting. A model was evaluated using the Boston housing dataset of 506 entries with 14 features. Comparing predicted to actual prices, the model acceptably predicted values within a margin of error of ±5. (Lancaster, 1966) and (Rosen, 1974) successively introduced the characteristic price model into the real estate industry. (Xia et al., 2020)adopted a logarithmic regression model to separately model whole second-hand houses, new houses under 5 years, and old houses over 5 years in Zhongyuan District, Zhengzhou.

In terms of the study on influencing factors of second-hand house price, (Butler, 1982) divided the influencing factors of housing price into architectural characteristics, neighborhood characteristics, and location characteristics, which were cited by many scholars in subsequent studies. (Dong et al., 2014) used second-hand and new house price indices from 16 cities including Beijing. Baidu search keywords were explanatory variables to build a model. Transactions, transaction taxes, and price trends were found important variables affecting housing price indices.

(Law, 2017) adopted a multi-level hedonic method, estimating the London region's impacts on housing prices. Differences in local streets were found to influence prices through impacts on travel. (Stadelmann, 2010) averaged housing prices in the Zurich metropolitan area using a Bayesian model with minimum variables as a priori constraints. Results showed in highly developed countries, real property characteristics of a location have a high probability of capitalizing on housing prices. (Hughes Jr and Sirmans, 1992) examined transportation's impact on housing prices. Despite traffic on all streets, the market adjusts residential prices according to high traffic externalities. (Churchill et al., 2021) used parametric and non-parametric panel methods to explore the nonlinear, time-varying link between transport infrastructure and housing prices. Results showed a positive correlation, with the nonparametric model revealing a changing relationship over time.

Existing home valuation often uses linear regression but with limiting assumptions. Recent studies apply machine learning techniques like regression trees, bagging, boosting, SVM, and random forests to better model price drivers. This study analyzes Wenzhou's existing home price factors. It collects 2799 data points on listings from Lianjia. Models are established using regression trees, bagging, boosting, SVM, and random forests, then compared via standardized error testing.

The main contributions of this paper are as follows: First, the second-hand house price estimation model of regression tree, bagging, lifting, support vector machine regression and random forest regression is established, and the standardized mean square error (NMSE) is used to test the model. Secondly, according to the random forest algorithm, the importance of the factors affecting the second-hand house price in Wenzhou is sorted. Third, explore the importance of different categories of second-hand housing price factors, broaden the scope of research on the importance of second-hand housing factors. Fourth, the regression tree analysis once again verified the importance of factors, and analyzed the marginal effect of different variables on the second-hand house price through the partial effect diagram.

2. Data source and variable processing

2.1. Data source

Homelinks is a leading existing home trading platform. This study extracts information on 2799 homes listed for sale in Wenzhou from Lianjia on March 18, 2020. As listings were from a single date, the data constitutes a cross-section, excluding price changes over time. Listings varied greatly by district - Lucheng had the most while Ouhai was second. Longwan, Yueqing City, and Ruian listings were almost equal. Cangnan, Pingyang, Dongtou, and Longgang had less information. No data was available for Taishun. Therefore, this study uses existing home price data from Lucheng, Ouhai, Longwan, Yueqing City, Ruian, and Yongjia - regions with the most comprehensive listings. This provides a representative sample to analyze factors influencing existing home prices in Wenzhou.

2.2. Variable selection

This study selects 11 variables as explanatory factors influencing second-hand house prices in Wenzhou: District, Number of Bedrooms, Number of Bathrooms, Total Floors, Floor Level, Floor Area, Property Orientation, Renovation, Near Public Transport, Number of High Schools within 1km, and Near Hospital. The listed unit price on the Lianjia website is the explained variable. Selection of explanatory variables: (1) Floor area affects prices within communities, with smaller areas having higher unit prices due to greater utilization. (2) Renovation impacts utility as rougher homes require decoration spending while finer offers immediate enjoyment, increasing willingness to pay. (3) Bedrooms determine occupancy and bathrooms impact convenience, directly affecting sale price. (4) South orientation receives stronger sunlight, impacting prices as buyers prefer south-facing homes. (5) Total floors indicate building age; older neighborhoods have fewer floors while higher imply newer construction, ideal facilities, and management potentially increasing prices. (6) Floor level balances convenience and benefits, with middle levels most desirable. (7) District differences in resources, salaries, and development impact prices as buyers select locations. (8) Nearby school numbers indicate educational resources, fetching premiums under limited enrollment. (9) Near transport and hospitals represent conveniences judged on maps.

2.3. Data processing

Raw data involved grading text-based attributes like relative floor, orientation, renovations, urban area, and proximity. Orientation facing south was assigned 1, others 0. Urban areas ranked by 2020 GDP. Relative floors coded as high=1, low=2, middle=3. Proximity to transport and hospitals used dummy variables (Table 1). Nearby school count within 1km directly used. Floor area, bedrooms, bathrooms, and floors used raw values. This grading/coding enabled qualitative variables' inclusion in quantitative models. Text attributes are standardized to facilitate regression/machine learning.

Table 1: Qualitative variable index quantification standard

Name of variable	Quantification
Orientation	Southeast, Southwest, Due South -1, other -0
Renovations Situation	Wool embryo -1, plain -2, hardcover -3
Part of Town	Yongjia County -1, Ouhai District -2, Longwan District -3, Ruian City -4, Lucheng District -5, Yueqing City -6
Whether nearby bus	Yes -1, no -0
Proximity to a hospital	Yes -1, no -0
Relative floor	High floor -1, Low floor -2, middle floor -3

2.4. Descriptive statistics

1. As can be seen from Figure 2 , housing price per square meter is right-skewed. Most homes cost around RMB 20,000/square meter. The mean is RMB 23,300/square meter while the median is RMB 21,800/square

meter, with some outliers as high as RMB 101,800/square meter. Nearly 60% of prices exceed RMB 20,000/square meter. In summary, prices generally cluster around RMB 20,000/square meter but outliers pull the mean up, so the median better represents the central tendency for this skewed data.

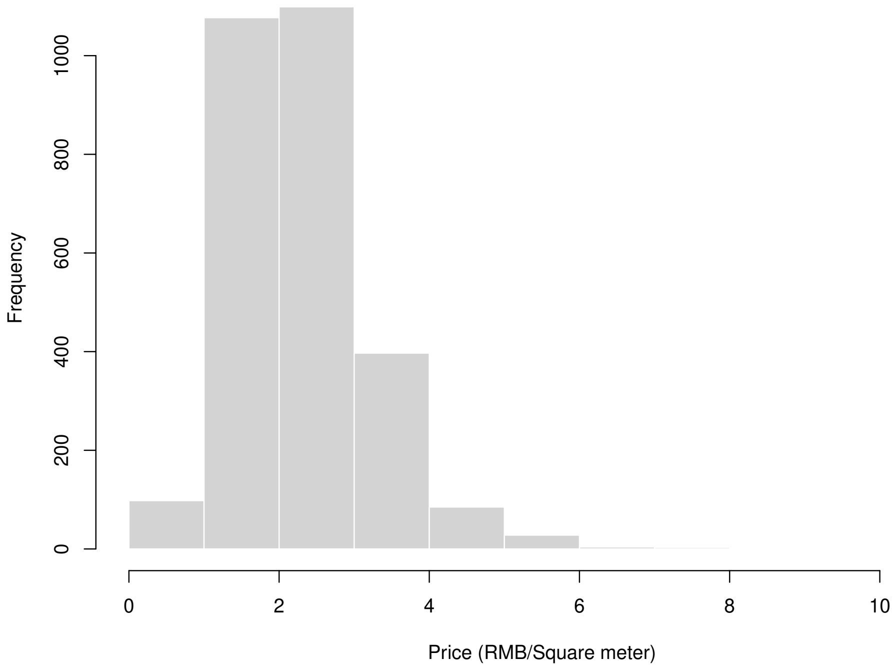

Figure 1: Price histogram of second-hand house per unit area

2. Unit Prices by Urban Area

As can be seen from Figure 2, the unit price of second-hand houses varies greatly among counties and cities. The six urban areas are divided into two echelons by unit area housing prices: the first echelon is Lucheng District, Ruian City and Ouhai District; The second echelon is Yongjia County, Longwan District and Yueqing City. Obviously, the secondhand house price per unit area of the first echelon is significantly higher than that of the second echelon, which may be due to the resource differences of each city, such as the convenience of transportation, the number of key primary and secondary schools and so on.

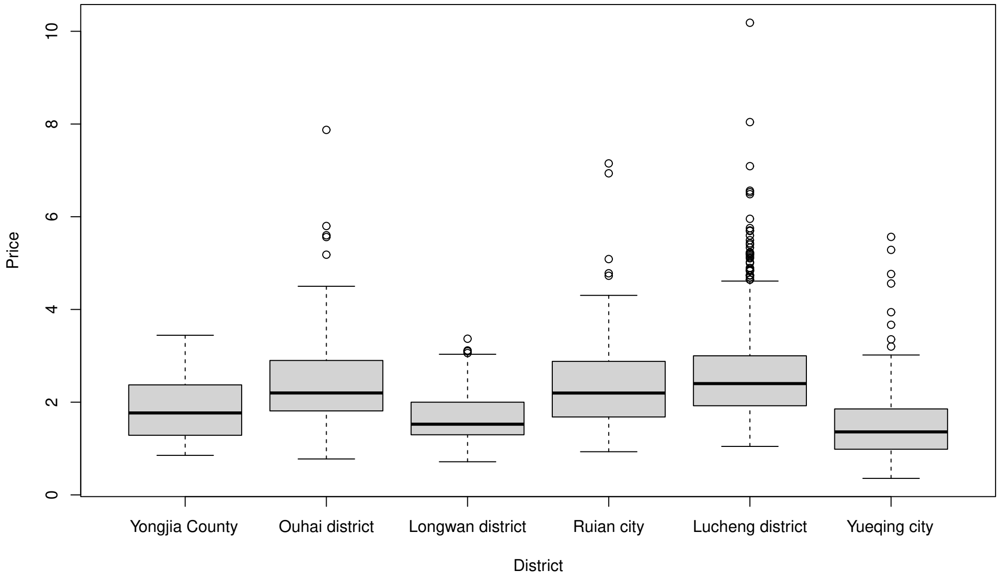

Figure 2: Distribution of housing price per unit area in each urban area

3. Impact of Bedrooms and Bathrooms

Figure 3 shows how bedrooms and bathrooms impact unit prices. For bedrooms (left), the price significantly differs by number, correlating higher with more bedrooms. For bathrooms (right), the price gradually increases with quantity. In summary, both variables determine price, though bedrooms exhibit a stronger relationship as more enable higher occupancy, substantially affecting unit cost. Bathrooms also influence pricing but to a lesser degree by enhancing convenience.

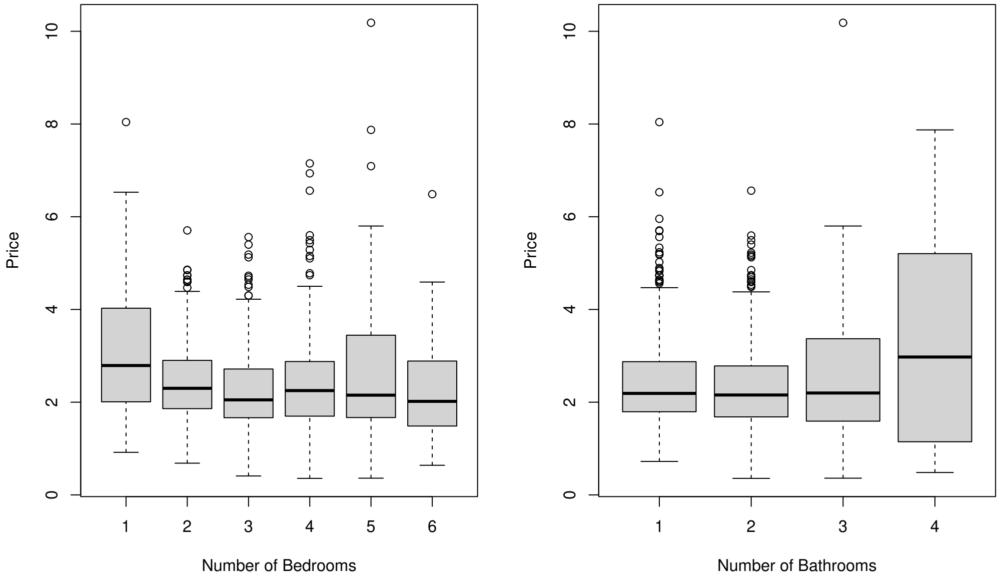

Figure 3: The distribution of house prices per unit area by number of bedrooms (left) and bathrooms (right)

4. Relative floor.

As shown in Figure 4, the unit area housing prices of low floors, middle floors, and high floors have little difference, and there are high abnormal points.

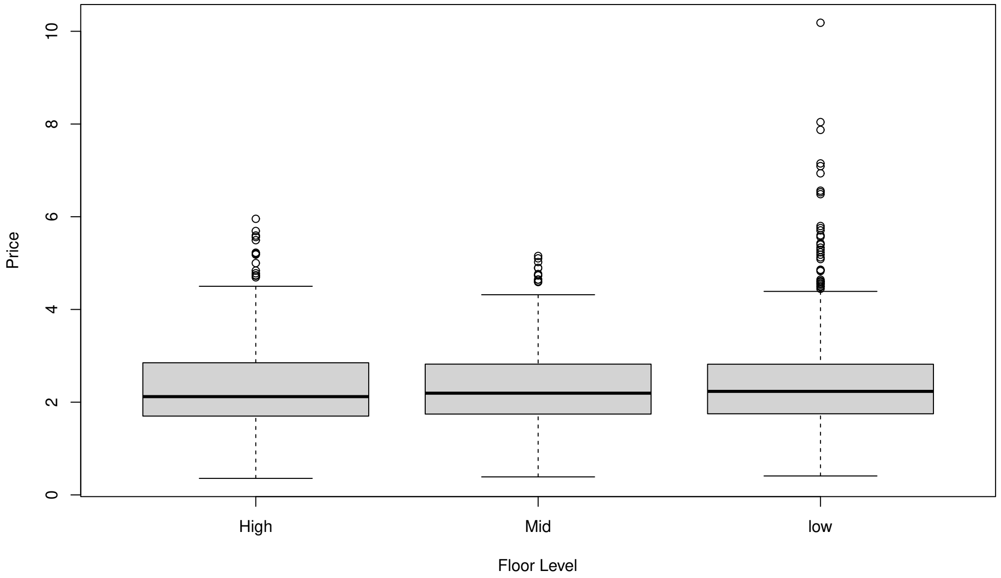

Figure 4: Price distribution map per unit area of different relative floors

3. Model comparison based on machine learning algorithm

3.1. Model introduction

Machine learning algorithms include regression tree, boosting, bagging, support vector machine and random forest algorithm. The following are introduced respectively: (1) Regression trees create a top-down tree structure by recursively allocating samples to child nodes based on feature threshold tests, enabling regression predictions. (2) Boosting combines weak learners into strong ones by iteratively reweighting samples, updating weights, and combining learners. (3) Bagging generates multiple training sets by sampling with replacement, training a base learner on each, and aggregating predictions, reducing variance versus a single learner. (4) Support vector machines minimize total loss within tolerance bands around a linear function to improve flexibility during training. (5) Random forests like bagging train multiple trees but consider a random feature subset at each node, averaging predictions for the result. Overall, these algorithms aim to combine multiple models to improve performance over a single base learner, through concepts like gradient boosting, averaging predictions, or focusing on varying subsets of features and training data.

3.2. Model comparison

This study first uses quantized data to construct and compare prediction models. Data is divided into training and test sets via 10-fold cross-validation for model fitting and evaluation. Normalized Mean Squared Error (NMSE) is calculated for regression tree, bagging, boosting, support vector machine, and random forest on both training and test sets. The model with the lowest average NMSE is identified as best fitting. Secondly, optimal model parameters are tuned to maximize accuracy. NMSE measures model fit, with lower values indicating better prediction performance. NMSE is defined as NMSE $= \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}$, Where y is the actual listed price and ŷ is the predicted value. NMSE values allow comparison of prediction ability across the five models (Table 2). Random forest exhibited the lowest average test set NMSE, demonstrating superior generalization performance. Regression tree, bagging, and SVM had similar fitting abilities. Boosting performed worst with the highest NMSE. While parameter tuning can improve individual model precision, random forest exhibited the strongest inherent generalization capacity for Wenzhou second-hand housing prices based on comparable model evaluations without external parameter influence. In summary, random forest regression is identified as the optimal predictive model for this application.

Table 2: Summary table of five model cross-validation results

Models	Training set mean accuracy (%)	Test set average accuracy (%)
Regression tree	67.8	71.8
Bagging	62.8	67.1
Boosting	73.9	74.1
Random forest	18.5	52.3
Support vector machine	63.3	72.6

4. Analysis of influencing factors of second-hand house price

4.1. Parameter determination of random forest algorithm

This study optimizes the random forest model for best fit and variable importance ranking. First, the number of decision trees (NTrees) is determined.

The number of variables (mtry) selected for each split must be specified initially. Prior research by (Handzel et al., 2012) found better performance when mtry=p/3, where p is the number of input variables. With 11 variables in this model, mtry was set to 4. Model error was plotted against increasing NTrees (Figure 5). Error decreases gradually as trees increase, stabilizing when NTrees surpasses 200. Therefore, NTrees=200 was selected for subsequent modeling based on Figure 5. At this point, increasing tree numbers provided diminishing returns with little further error reduction. In summary, an empirical approach was taken to optimize the random forest hyperparameters based on established guidelines and the evaluation of model error against tree count. This tuned model provides the basis for understanding variable importance using random forest techniques to predict second-hand housing prices.

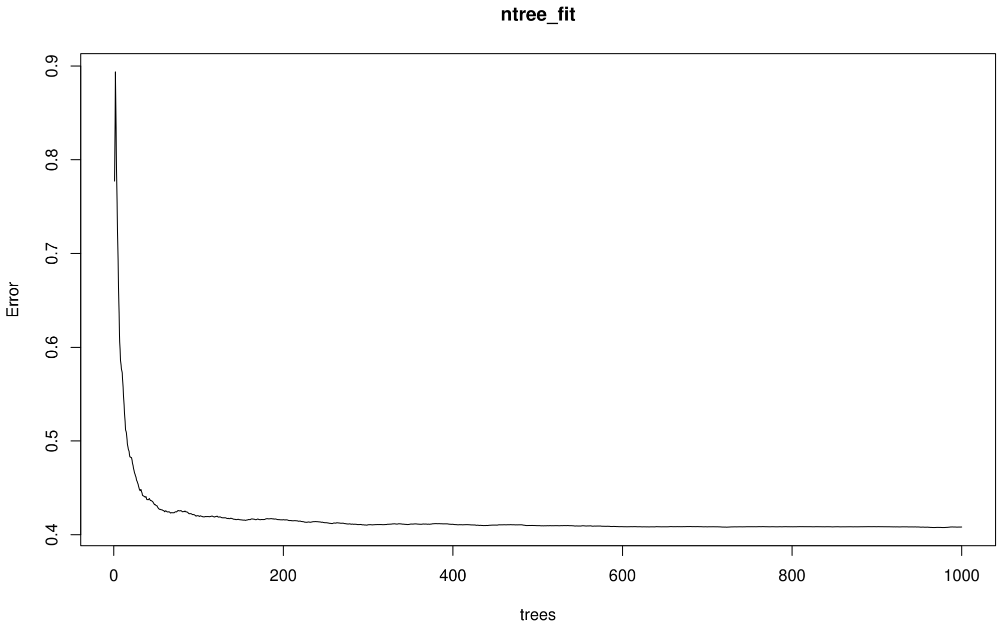

Figure 5: Error variation with the number of decision trees

Secondly, the optimal number of randomly sampled variables (mtry) per split was determined. mtry represents variables considered at each node in random forest models. Figure 6 evaluates model fit across mtry 1-10, with ntree at 200. The residual sum of squares decreases and goodness of fit increases with rising mtry, leveling off beyond mtry=5. Therefore, mtry=5 was determined to optimize variable candidates for split consideration, achieving model fit without excess improvement above this threshold. In summary, tuning mtry alongside ntree helped maximize random forest performance for predicting second-hand housing prices.

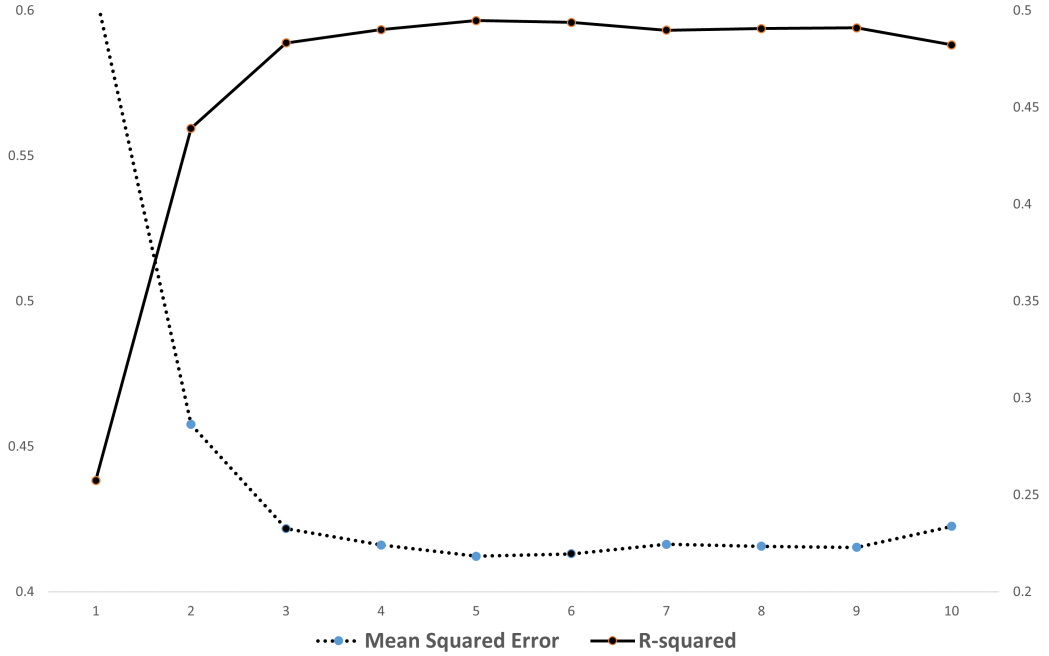

Figure 6: Comparison diagram of different mtry models

4.2. Importance analysis of influencing factors

The model used optimized hyperparameters ntree=200, mtry=5. The average residual sum of squares was 0.412, indicating a good regression fit. Variable importance was measured via IncMSE (variable noise sensitivity) and IncNodePurity (model fit influence). Table 3 shows the results. The top 5 most important by IncMSE were: urban area, total floors, floor area, nearby school count, and renovation level. Top by IncNodePurity were: floor area, total floors, urban area, nearby school count, and bathroom count. In summary, urban area, floor area, total floors, and nearby school count emerged as having the strongest influence on prices overall per this random forest model, providing insight into key price determinants. The tuned algorithm and metrics identified important variables.

Table 3: Data table of the importance of factors affecting the price of second-hand housing

Variable	IncMSE(%)	IncNodePurity(%)
District	83.161	376.955
Number of Bedrooms	16.115	106.478
Number of Bathrooms	20.369	136.272
Total Floors	64.789	513.416
Floor Level	9.222	86.982
Floor Area	35.788	518.873
Property Orientation	5.358	12.305
Renovation	25.164	105.481
Near Public Transport	20.491	49.014
Number of High Schools within 1km	26.483	150.705
Near Hospital	-2.258	23.925

4.3. Echelon division of influencing factors

As shown in Figure 7, according to the increase of mean square error as the reference standard, the importance of variables can be sorted into three echelons according to the importance level: Tier 1 (most important): Urban area, total floors, floor area, nearby primary/secondary schools; Tier 2: Renovation level, proximity to bus, bathroom count; Tier 3 (least important): Bedroom count, floor level, orientation, proximity to hospital; The first tier factors - urban area, total floors, floor area, and nearby school count - most influenced Wenzhou prices. Urban area differences in transport, education, healthcare, and development drive buyer preferences. Total floors signify newer construction and better amenities/management. Floor area directly relates to livability, with a large price effect. School count importance stems from Wenzhou's enrollment policies. Third-tier factors like bedrooms, floor level, orientation, and hospitals had minimal impact as standard designs address these. Housing mainly meets living needs, with bedrooms in a typical 2-3 range seeing diminishing returns. Elevators mitigate floor impacts. Developments include south orientations. Hospital proximity is less priority. In conclusion, location and development attributes linked to livability, and investment potential conferred the strongest pricing influence, underscoring buyers' key concerns. Microproperty details mattered less.

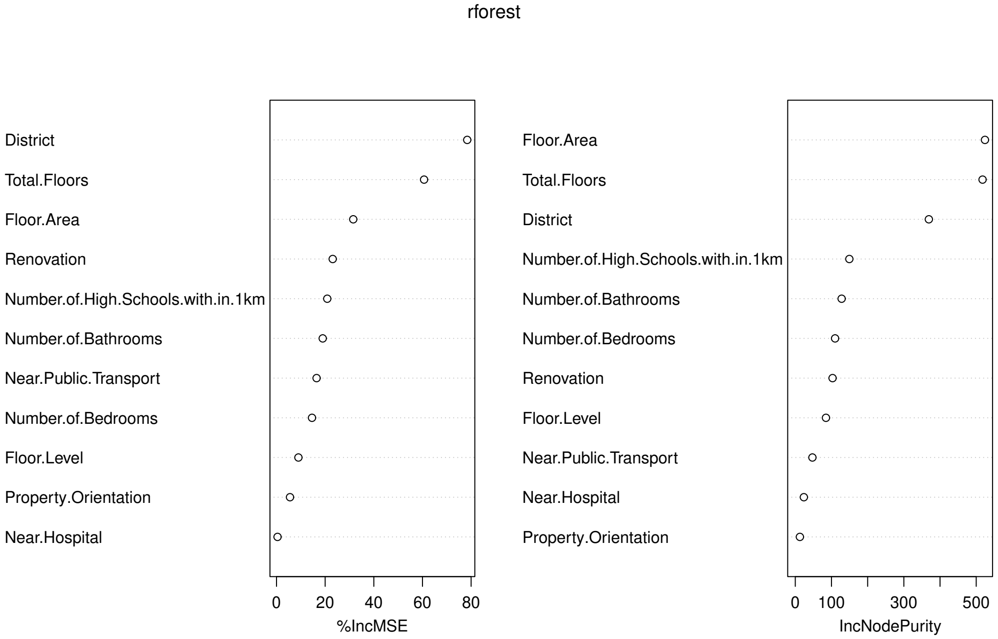

Figure 7: Ranking chart of the importance of factors affecting the price of second-hand housing

5. Heterogeneity test

5.1. Heterogeneity analysis of the urban area

This study examined factors influencing second-hand housing prices across six Wenzhou urban areas. Total floors and floor area were commonly important across areas. In Yongjia County, the top factors were: total floors, renovation, bedrooms, floor area, proximity to transit. For Ouhai District: total floors, bathrooms, transit access, floor area, renovation. In Longwan District: total floors, floor area, bedrooms, renovation, nearby schools. In Ruian City: renovation, total floors, floor area, transit access, bathrooms. For Lucheng District: floor area, total floors, nearby schools, transit access, bathrooms. In Yueqing: total floors, floor area, nearby schools, renovation, bedrooms. Renovation had a larger effect in Yongjia County and Ruian City likely due to lower upgraded supply stimulating higher demand. Nearby transit influenced prices more in areas with well-developed public transportation systems. Nearby school availability mattered more where academic competition is intense. Bathroom counts influenced Ouhai District more. Bedrooms counted more in Yongjia County and Longwan District due to typically larger household sizes. In conclusion, while some factors generally determined prices, demand priorities explained local value variations. Pricing responded to meeting area-specific livability needs.

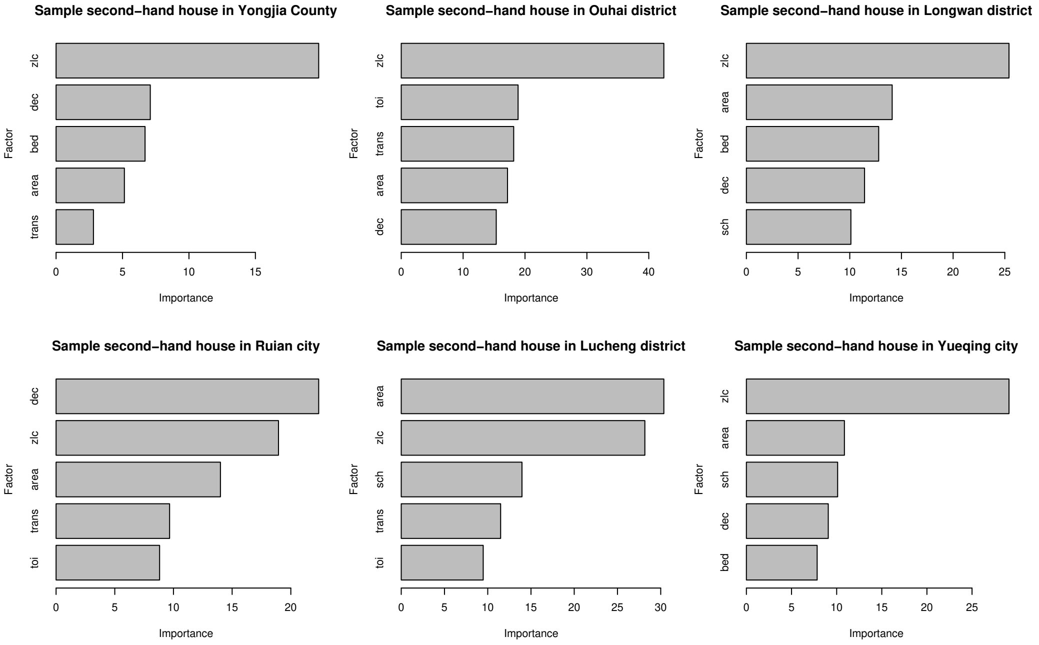

Figure 8: Heterogeneity analysis based on urban area

5.2. Heterogeneity analysis of total floors

This study examined how building height influenced important second-hand housing factors. Properties were categorized as: low-rise (1-3 floors), mid-rise (4-6 floors), mid-high rise (7-9 floors), and high-rise (10+ floors). As Figure 9 shows, for low-rise homes, bathrooms, floor area, city, bedrooms, and nearby schools most impacted price. For mid-rise homes, influential factors were city, floor area, bathrooms, renovations, and bedrooms. Mid-high rise property values responded to city, floor area, bedrooms, bathrooms, and transit access. High-rise home prices were driven by city, floor area, nearby schools, renovations, and transit access. As high-rises comprised 59.7% of samples, their drivers aligned with overall patterns. However, bathroom and bedroom counts more strongly shaped prices for low-rise, mid-rise and mid-high rise properties. This effect was most pronounced for bathrooms in low-rise homes, exceeding impacts of floor area and city. This likely reflects larger household sizes typically associated with older, self-built low-rise properties versus newer high-rises housing smaller family units. Additional bathrooms and bedrooms better accommodated larger households' needs. In summary, while certain drivers were common, demand priorities explained inter-height variations in factor importance. Pricing responded to meeting buyers' lifestyle needs given property characteristics.

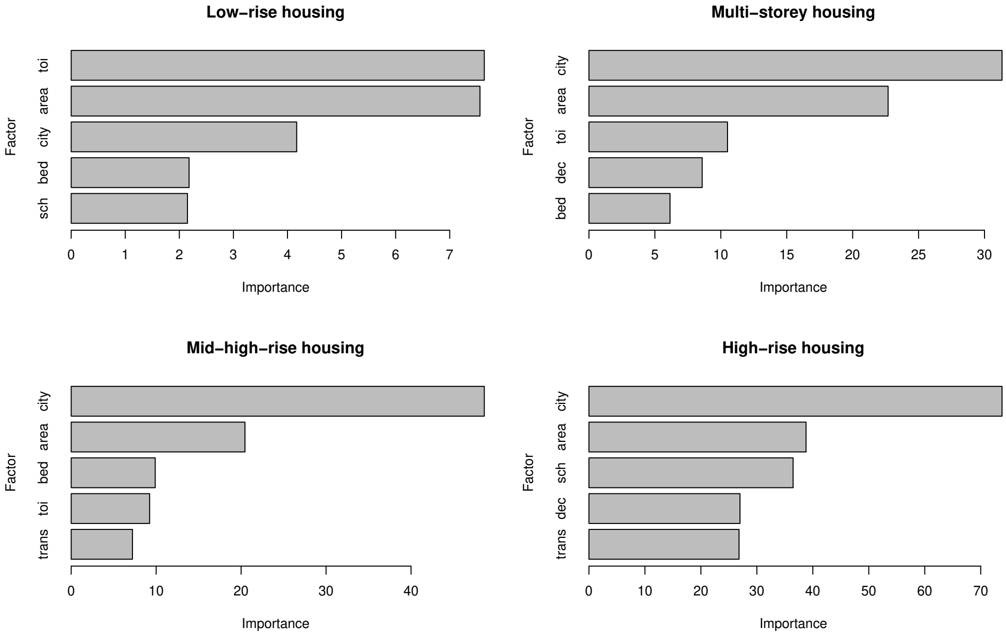

Figure 9: Heterogeneity analysis based on total floors

5.3. Heterogeneity analysis of building area

This study examined how housing size influenced the factors determining second-hand home prices. Properties were categorized as small (<90 sqm), medium (90-144 sqm), or large (>144 sqm). As shown in Figure 10, for small homes the top five factors were: city, total floors, bedrooms, nearby schools, and bathrooms. For medium homes, the top factors were: city, total floors, renovations, nearby schools, and proximity to transit. The top large home determinants were: city, total floors, renovations, bathrooms, and nearby schools. Across sizes, city and total floors remained important. However, renovations carried more weight for medium and large homes likely due to higher renovation costs relative to space. Small home renovations require less time and expense. Bedrooms mattered more for small homes, which have constrained space, whereas medium and large layouts can accommodate families without major bedroom impacts on price. In summary, while location and building attributes fundamentally drive pricing, floor area subtleties like renovation feasibility and bedroom optimization influence which characteristics have amplified roles. Demand priorities thereby differ by property size.

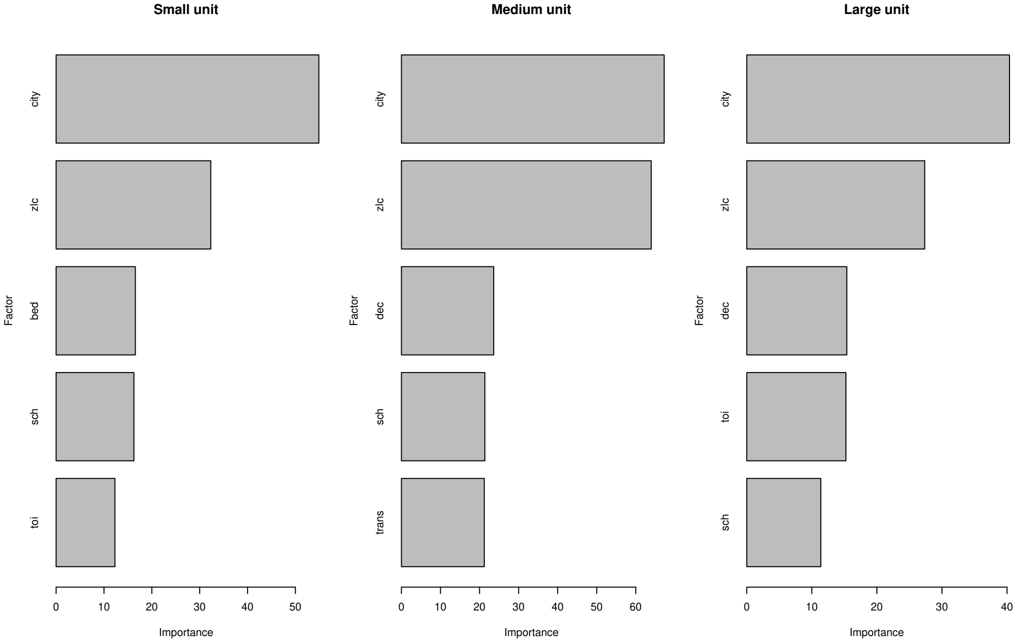

Figure 10: Heterogeneity analysis based on building area

6. Optimized path analysis

6.1. Regression tree visualization

Figure 11 displays a regression tree from random forest modeling, which splits nodes based on the greatest impurity decrease. Higher nodes influence more data. The tree shows the urban area, total floors, and floor area as top-splitting variables, consistent with importance metrics. This confirms regression tree structure aligns with random forest results, validating empirical conclusions. Examining tree structure demonstrates variable prioritization logic within modeling. Both importance outputs and tree visualization identify urban areas, total floors, and floor areas as primary price drivers in Wenzhou. The replication of findings increases confidence in the random forest approach.

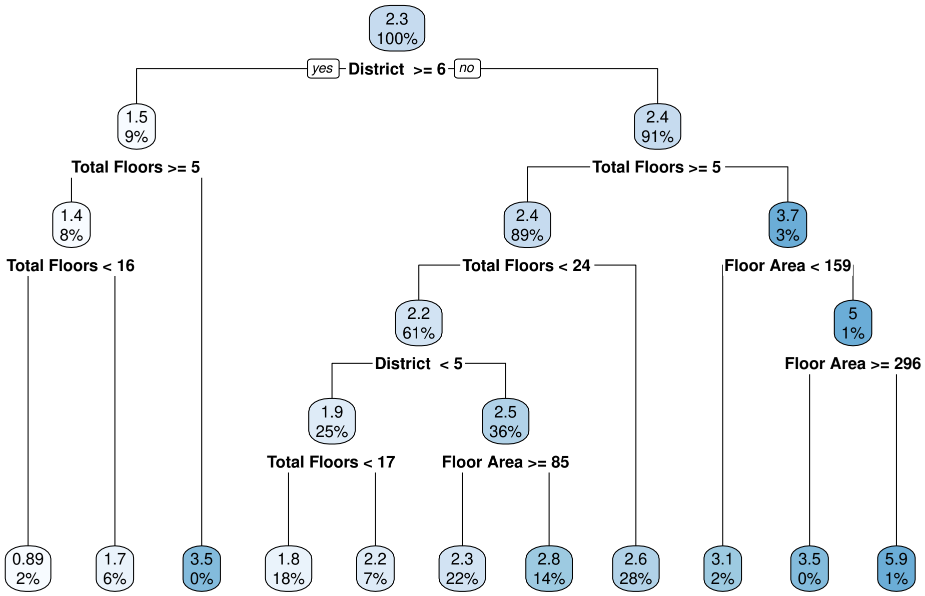

Figure 11: Visual result of regression tree

6.2. Bias effect graph analysis

Figure 12 shows marginal effects plots of important variables on price with other attributes held constant. The plots illustrate non-constant marginal impacts as total floors, floor area, nearby schools, and bathrooms vary, exhibited by multiple inflection points on each curve. Rather than uniform effects, certain intervals correspond to greater incremental price impacts for a given characteristic than others. These marginal effects plots provide added insight beyond average effects by revealing varying price sensitivity within each variable's distribution.

The marginal effects plots provide insights into how the variables differently impact prices at varying levels: Firstly, total floors showed the steepest slope from 1-10, with the price lowest at 10 floors. This indicates the marginal effect on price was greatest as floors increased from 1 to 10, though the impact was negative. Secondly, floor area had the maximum negative slope from 20-60 sqm, suggesting price declined most rapidly within this range. Smaller units turnover quickly not for living, inflating their price relative to satisfying housing needs. Thirdly, the nearby school count was most influential rising from 0 to 1, changing an area from a non-school to a school district. Additional schools only modestly impacted prices thereafter by broadening options. Finally, bathrooms exhibited the steepest positive slope from 2-4 bathrooms. Demand for and pricing power of additional bathrooms likely peaks within this common configuration, before marginal returns diminish with further additions exceeding typical needs. In summary, the marginal effects plots illustrate how price responsiveness varies non-linearly across different portions of each variable's distribution, shedding light on demand determinants beyond average effects alone.

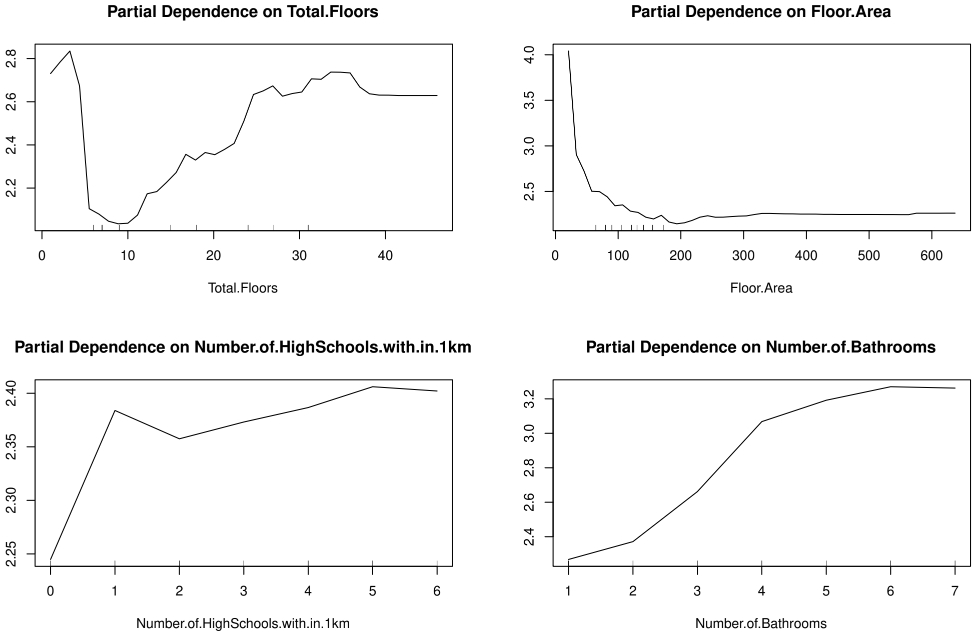

Figure 12: Partial effect diagram of the total floor, building area, number of primary and secondary schools nearby, and number of toilets

7. Conclusion and suggestion

7.1. Conclusion

This study collected sales data for 2,799 second-hand homes in Wenzhou from an online real estate platform. Data included 11 influential factors. Five regression models were established and compared: regression trees, bagging, boosting, random forest, and support vector machines. Cross-validation, heterogeneity testing, and marginal effects analysis yielded the following conclusions:

(1) Models were trained and tested on the same datasets to intuitively compare performance. Random forest had the lowest NMSE on test data, indicating the strongest generalizability. Boosting performed worst with the highest NMSE.

(2) The optimal random forest parameters for predicting prices were: 200 trees and five variables sampled at each split. The model fits well with 0.412 residual sum of squares. Variable importance ranked: urban area, total floors, floor area, and nearby schools as most important.

(3) The importance of factors varied by urban area. Total floors and floor area universally impacted prices. Renovation mattered more in some cities. Proximity to transport influenced prices where infrastructure was highly developed. Nearby schools impacted areas with competitive education.

(4) Marginal effects differed across variable ranges. Total floors from 1-10 and floor areas from 20-60 sqm had strongly negative impacts. The nearby school counts from 0 to 1 conferred large positive changes. Bathroom counts from 2-4 conveyed greater positive impacts.

In conclusion, random forest best-predicted prices. Variable importance depended on property and location attributes. Marginal effects revealed nonlinear relationships between factors and pricing.

7.2. Suggestion

Based on the above research conclusions, the following suggestions are put forward:

(1) For government: Strengthen supervision of secondary housing platforms to curb malpractice. Prices clearly relate to region-specific attributes; tailored policies recognizing inter-area differences are needed. For example, improving public transport access in areas where this influences prices. Also, stabilize prices while steadily raising incomes to reasonably fulfill housing's core living function.

(2) For buyers and sellers: Sellers can assess property value according to important factor influences identified, comparatively evaluating listed prices to avoid information asymmetry. Buyers should reference factor importance degrees matching needs to circumspectly select homes avoiding undue economic losses. Promoting rational, informed transactions on both sides requires understanding secondary market drivers. Sellers gain pricing power insights; buyers uncover what truly impacts affordability satisfaction. Government oversight combined with educated market engagement better serves housing security objectives for all.

In sum, this research untangles critical purchase determinants. Applying learnings can help balancing stakeholder priorities through responsible development, balanced policies and prudent consumer decision-making. An evidence-based approach supports sustainable secondary markets.

Appendix A. Funding agency

The research was partially supported in part by the National Natural Science Foundation of China with award number 62372409, by the Ministry of Science and Technology with award number DL2023147001L, and by the Zhejiang University with award number 226202200238.

References

Adetunji, A.B., Akande, O.N., Ajala, F.A., Oyewo, O., Akande, Y.F., Oluwadara, G., 2022. House price prediction using random forest machine learning technique. Procedia Computer Science 199, 806–813.

Afonso, B., Melo, L., Oliveira, W., Sousa, S., Berton, L., 2019. Housing prices prediction with a deep learning and random forest ensemble, in: Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional, SBC. pp. 389–400.

Antipov, E.A., Pokryshevskaya, E.B., 2012. Mass appraisal of residential apartments: An application of random forest for valuation and a cart-based approach for model diagnostics. Expert Systems with Applications 39, 1772–1778.

Butler, R.V., 1982. The specification of hedonic indexes for urban housing. Land economics 58, 96–108.

Churchill, S.A., Baako, K.T., Mintah, K., Zhang, Q., 2021. Transport infrastructure and house prices in the long run. Transport Policy 112, 1–12.

Dong, Q., Sun, N., Li, W., 2014. Real estate price prediction based on search data. Statistical Research , 8.

Foryś, I., 2022. Machine learning in house price analysis: regression models versus neural networks. Procedia Computer Science 207, 435–445.

Handzel, A.A., Liaw, A., Stefano, F.P., Chan, G.K., Szewczak, A.A., Santini, F., 2012. Data analysis approaches for high content screening. Statistics in Biopharmaceutical Research 4, 194–204.

Ho, W.K., Tang, B.S., Wong, S.W., 2021. Predicting property prices with machine learning algorithms. Journal of Property Research 38, 48–70.

Hughes Jr, W.T., Sirmans, C., 1992. Traffic externalities and single-family house prices. Journal of regional science 32, 487–500.

Lahmiri, S., Bekiros, S., Avdoulas, C., 2023. A comparative assessment of machine learning methods for predicting housing prices using bayesian optimization. Decision Analytics Journal 6, 100166.

Lancaster, K.J., 1966. A new approach to consumer theory. Journal of political economy 74, 132–157.

Law, S., 2017. Defining street-based local area and measuring its effect on house price using a hedonic price approach: The case study of metropolitan london. Cities 60, 166–179.

Park, B., Bae, J.K., 2015. Using machine learning algorithms for housing price prediction: The case of fairfax county, virginia housing data. Expert systems with applications 42, 2928–2934.

Rosen, S., 1974. Hedonic prices and implicit markets: product differentiation in pure competition. Journal of political economy 82, 34–55.

Stadelmann, D., 2010. Which factors capitalize into house prices? a bayesian averaging approach. Journal of Housing Economics 19, 180–204.

Xia, Q., Lu, J., Liu, C., Jia, X., 2020. Study on hedonic price of second-hand housing in zhongyuan district of zhengzhou city under the background of large data. Areal research and development 39, 6.