DOI: 10.54254/2753-8818/2026.DL34409

Feature Attribution and Error Diagnosis of a Random Forest Housing Price Prediction Model Integrating SHAP and LIME

Ruoxian Wang

Institute of Science, Minzu University of China, Beijing, China
chordwangruoxian@outlook.com

Abstract. Housing is one of the most basic needs of humanity globally, and fluctuations in housing prices have garnered significant attention. Currently, the scale of China's second-hand housing transaction market is rapidly expanding. Understanding the factors influencing second-hand housing prices can enhance transaction efficiency and achieve a balance between supply and demand. This article obtains the Ames dataset from Kaggle and selects the Random Forest (RF) model, which is a relatively mature model in the field of housing price prediction, for training. It integrates Shapley Additive exPlanations (SHAP) and the Local Interpretable Model-agnostic Explanation (LIME) to analyze the importance of housing features. The results indicate that the RF model exhibits good performance in predicting housing prices. The SHAP analysis reveals that the most significant global features influencing second-hand housing prices are the quality of building materials and decoration, followed by living area, and there is an interaction between these two factors. Furthermore, it utilizes LIME to perform local attribution on high-quality small-area samples in the test set. The results revealed that the local contribution of living area in such samples was significantly reduced, and replaced by features such as housing functionality, indicating that the decision logic of the RF model may deviate in local areas that deviate from the mainstream distribution.

Keywords: second-hand housing prices, random forest, SHAP method, LIME, price forecast

1. Introduction

In the research on the influencing factors of second-hand housing prices, traditional methods mostly involve constructing multiple linear regression models for analysis. However, the real estate market is a complex nonlinear system, and traditional linear models are sensitive to outliers in data. House price data with skewed distributions can affect the robustness of the model.

Currently, the new generation of artificial intelligence, as a digital tool, is thriving and has become an effective way to enhance the accuracy of real estate appraisal. Based on machine learning and modeling, Hong complements traditional regression methods by studying the linear or nonlinear relationship between dependent and independent variables, as well as the hierarchical structure of factors determining housing prices [1]. Shi conducted training and prediction of second-hand

DOI: 10.54254/2753-8818/2026.DL34409

housing prices based on five different models, and found that the final fitting effect was best for Random Forest and worst for Lasso regression [2]. When Shi was conducting high-precision real estate value prediction, he found that under the same experimental environment, the prediction performance of the random forest model was significantly better than that of the deep neural network [3]. After integrating XGBoost with the RF model for housing price prediction, Zhao found that the prediction accuracy was significantly higher than that of either model alone [4]. Based on the second-hand housing data from Homelink, Du utilized the SHAP method for both global and local explanations, quantitatively assessing the decision-making weight of each feature within different models [5]. This article will integrate the comprehensive findings from the aforementioned research and select an appropriate model for housing price prediction and error diagnosis.

This paper will take the Ames data set as an example, carry out multi-dimensional analysis on the data of second-hand housing through RF, and explain the model afterwards with the methods of shap and lime. The purpose is to deeply explore the key features affecting second-hand house prices and verify the local accuracy of RF and Shap methods.

2. Methods

2.1. Data source and description

The data used in this study is the Ames housing dataset on Kaggle. The data set is from the evaluation office in Ames, Iowa, USA, and covers the information of residential properties sold from 2006 to 2010 [6]. The original data set contains 2930 observation records, involving 82 variables, including 23 nominal variables, 23 ordered variables, 14 discrete variables, and 20 continuous variables. The target variable is Sale Price. The Ames data set can be used to calculate the value of real estate appraisal, as well as the physical measurement indicators of real estate and the computational variables used in the city's real estate appraisal process [6]. Compared with the classical Boston housing data, this sample size is larger, and the variables are richer.

2.2. Index selection and description

The explanatory variable of this study is the housing price, which is taken as the natural logarithm before modeling, and its influencing factors are studied from the four dimensions of building structure, location characteristics, indoor facilities, and external environment.

First, preprocess the data. Eliminate variables with a missing rate of more than 90%, and process the missing value according to the meaning of the variable. Fill none for category variables without and 0 for numerical variables; Missing and median are filled in for the remaining categories and missing numerical variables, respectively.

In order to ensure that the random forest can handle numerical input, the classification variable is transformed into numerical form in this study. The ordered variables are mapped to integers from 5 to 1 according to the level, and the nominal variables are encoded by the unique hot code. In addition, new features are constructed: the age of houses is calculated by year of construction and year of sale; Convert continuous variables such as pool area and fireplace area into binary variables (yes/no).

DOI: 10.54254/2753-8818/2026.DL34409

2.3. Method introduction

RF is an integrated learning method based on a decision tree. RF gives the results of competition with boosting and adaptive bagging, but does not gradually change the training set. The accuracy of the processing results shows their ability to reduce the deviation [7]. In this study, the model parameters were adjusted by cross-validation and evaluated by mean square error (MSE), root mean square error (RMSE), and coefficient of determination (R^2).

SHAP is a method based on the Shapley value in cooperative game theory to evaluate and explain the feature importance of a machine learning model. It decomposes the predicted value of the model into the sum of the contribution values of each feature, which can reflect the global feature contribution and local feature contribution of the model prediction at the same time, and the visualization effect is excellent [8].

For the tree model, the TreeSHAP algorithm can use the tree structure to efficiently calculate the accurate Shap value, which meets the local accuracy. When a feature is missing, the shape value can be directly set to 0 to avoid filling the introduced deviation.

LIME is an interpretation method for the prediction results of a single sample to verify the prediction accuracy of the model in a local range [9]. A simple interpretable surrogate model is used to approximate the local decision boundary of the original black box model [10]. In this study, sparse linear regression is used as a surrogate model to approximate the local decision boundary of RF, so as to obtain the local characteristic attribution of samples and display the characteristic interval and its corresponding local linear weight.

LIME only makes a linear approximation around the sample and is not biased by the global average behavior. It can provide concise and focused important features in the face of high-dimensional data.

3. Results and discussion

3.1. RF training results

Firstly, train the train dataset in the Ames dataset and construct an RF model. After five-fold cross-validation, the performance of the model is presented in Table 1:

Table 1. RF performance verification

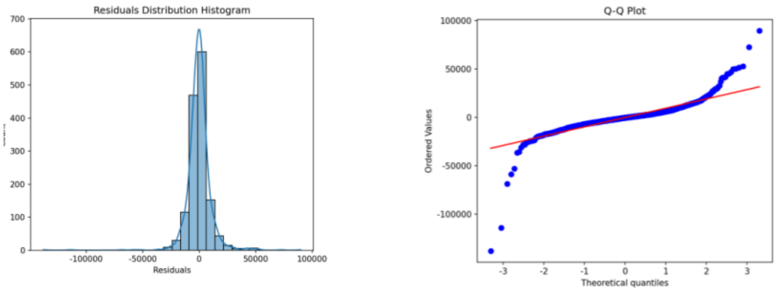
Metric	R^2 Score	MSE	RMSE (USD)	MAE	MAPE
Value	0.84 ± 0.10	$9.77\times 10^8 \pm 5.41\times 10^8$	30289.59±7698.85	17811.77±1886.38	10.43%±1.26%

This model can explain over 80% of the variance in housing price fluctuations and exhibits good predictive ability. While there is some fluctuation in performance across different discounts, it remains stable overall. After five-fold cross-validation, the average error between the housing prices obtained from the RF model and the actual housing prices is around $30,000. The housing prices in the Ames dataset range from approximately $30,000 to $750,000, so this error is within a reasonable range.

3.2. RF model goodness of fit and residual diagnostics

To evaluate the applicability of RF for housing price prediction, this study plotted a histogram (Figure 1(a)) and a QQ plot (Figure 1(b)) of the residual distribution in the training set.

DOI: 10.54254/2753-8818/2026.DL34409



(a) (b)

Figure 1. (a)Residuals distribution histogram; (b)QQ plot (photo/picture credit: original)

As can be seen from Figure 1 (a), the residuals are symmetrically distributed around zero, the peak value is near zero, and the residuals range from about -150000 to 150000 dollars. The overall shape of the histogram is similar to the normal distribution, indicating that the residuals have good normality. Most of the points in Figure 1 (b) are closely distributed near the diagonal, especially in the middle segment. On the whole, there is no obvious S-shaped or systematic bending of the points in the QQ diagram, and the assumption of normality of the residuals is basically satisfied.

The good residual normality confirms the randomness and unbiasedness of the prediction error of the RF model, and the prediction accuracy of the model remains relatively stable in different housing price ranges. This provides a reliable model foundation for subsequent feature impact interpretation using SHAP and LIME methods.

3.3. SHAP interpretation and result visualization

SHAP calculates the SHAP value based on 500 pieces of data in the training set, balances efficiency and representativeness, and then generates a variety of graphs and outputs a list of average absolute SHAP values. The whole process aims to improve the transparency and interpretability of the model.

3.3.1. Bee swarm diagram

Figure 2 is a bee swarm diagram, showing the influence direction and intensity of each feature on the model output. Red represents high quality and blue represents low quality.

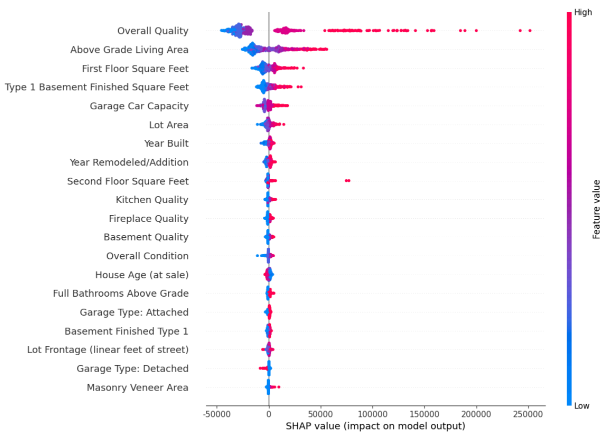


Figure 2. Bee swarm (photo/picture credit: original)

DOI: 10.54254/2753-8818/2026.DL34409

As can be seen from Figure 2, overall quality (OverallQual) is the most important feature affecting second-hand house prices: high quality can significantly increase house prices, while low quality can also significantly lower house prices. The ground living area (GrLiveArea)is the second important influencing factor, making a positive contribution in a large area. Fireplace quality (FirePlaceQu), kitchen quality (KitchenQual), and other quality features also show that high quality improves house prices. The height of the eigenvalue is consistent with the direction of the SHAP value.

3.3.2. Scatter plot of above-ground living area and first-floor area

Figure 3 shows the relationship between GrLiveArea and first floor area (1stFlrSF) and the shape value. The color represents overall quality, red represents high overall quality, and blue represents low overall quality.

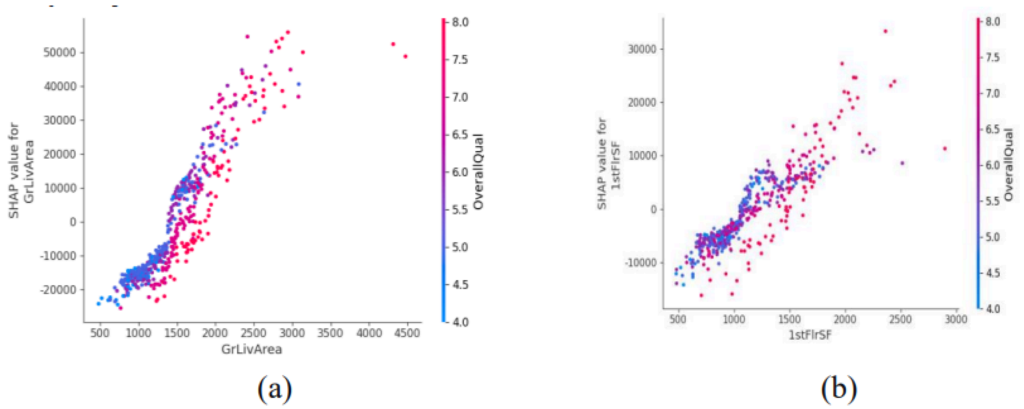


Figure 3. Scatter plot (a) GrLiveArea dependence; (b) 1stFlrSF dependence (photo/picture credit: original)

It can be seen from Figure 3 (a) that the above-ground living area has a non-linear positive correlation with the SHAP value: when the area is small, the SHAP value is negative, which turns positive with the increase of the area, but the marginal effect decreases. Figure 3 (b) shows that when the area of the first floor is less than 1200 square feet, the SHAP value is mostly negative or a small positive. After exceeding 1200, the shape value rises rapidly, and the higher the quality, the steeper the rise.

Figure 3 shows that there is an interactive relationship between quality and area. For second-hand houses, it is necessary to expand the area of houses on the basis of ensuring the quality of houses in order to give full play to the advantages of large units.

3.4. LIME analysis results

This study used live to screen the data of the test set with the overall quality greater than 7 points and the above-ground area less than 1300 square feet, and finally obtained 59 qualified samples of high-quality small-area second-hand houses. Taking the sample id=1465 as an example, the importance of its local characteristics is analyzed, as shown in Table 2.

DOI: 10.54254/2753-8818/2026.DL34409

Table 2. Top ten local feature weights with id=1465

Rank	Feature Range	Value
1	OverallQual > 7.00	83,033.3733
2	Home Functionality > 6.00	-20,633.1676
3	1129.50 < GrLivArea ≤ 1464.00	-17,433.5721
4	Kitchens Above Ground ≤ 1.00	7,753.2565
5	Lot Area ≤ 7553.50	-6,147.6120
6	1973.00 < Year Built ≤ 2000.00	5,863.1602
7	Fireplace Quality ≤ 0.00	-4,041.5724
8	1.00 < Garage Car Capacity ≤ 2.00	3,510.1748
9	Miscellaneous Value ≤ 0.00	-3,390.1735
10	Overall Condition ≤ 5.00	-3,288.1208

The overall quality of the house with test sample id=1465 is 8 points, the above-ground living area is 1280 square feet, and the RF predicted price is $189698.Among the top 10 importance of its life characteristics, the largest positive contribution is QverallQual>7, followed by home functionality>6. Further research found that the lot area of this sample was only 5005 and located in an unknown community. These negative characteristics were given a higher penalty weight in the local linear model. Therefore, the positive contribution of the large area GrLivArea was diluted.

The GrLivArea lime weights of the above 59 samples were analyzed in batches with lime, and the results showed that their average GrLivArea weight was -20441.84.

3.5. Comparative analysis of SHAP and LIME

The SHAP scatter plot showed that OverallQual had a positive regulatory effect on GrLivArea. However, the interpretation results of the selected counterexample samples by LIME show the opposite local pattern. The result of LIME shows that although the quality of the house is high, the area does not bring the expected high premium. This shows that the interaction effect of area and mass has local instability, and it will significantly attenuate or even reverse in atypical samples.

To sum up, SHAP provides the global attribution of numerical features, and lime outputs the local feature interpretation, which successfully identifies the deviation of the global conclusion of SHAP from the local atypical samples and the decision-making blind spot of the model. The integration of shape and life for feature attribution realizes a complete interpretable analysis framework from macro rules to micro individuals.

3.6. Limitations

Although the RF model used in this paper and the analysis method of integrating Shap and life have certain feasibility in predicting the price of second-hand housing, there are also some shortcomings.

First of all, because the data in the Kaggle database is used, the influencing factors of house prices are different from those in China. For example, the policy of single school partition and Nearby Enrollment implemented in China has led to the rise of school district house prices. Educational resources not only affect the second-hand houses in the area, but also affect the prices of new houses [11]. Secondly, this paper mainly uses machine learning algorithms. There are more methods to try in the problem of house price prediction. For example, Yang uses the XGBoost model

DOI: 10.54254/2753-8818/2026.DL34409

optimized by the sparrow search algorithm (SSA) to predict house prices. Compared with the traditional RF model, MAE is significantly reduced, and the problem of parameter adjustment is solved through automatic optimization [12]. It can also study the factors affecting house prices from a multidimensional perspective. For example, Liang uses the multi-scale geographically weighted regression model (mgwr) to capture the differences in the spatial action scale of the factors affecting house prices [13].

To sum up, future research can combine China's national conditions, select variables from different angles, optimize research methods, and conduct a deeper analysis of housing prices.

4. Conclusion

This study uses RF to predict Ames' second-hand house prices and discusses the key factors affecting second-hand house prices. The results show that the RF model has a good prediction ability with an average R^2 of 0.8425 after 50% cross validation. The overall material and decoration quality of the house are the core factors affecting the house price; The second is the above-ground living area.SHAP further revealed that there was a significant synergistic effect between the overall quality and residential area. LIME's analysis of high-quality, small-area second-hand houses found that the impact of aboveground living area on house prices was significantly reduced. Therefore, it is necessary to carry out local diagnosis for atypical samples while carrying out global explanatory analysis on the black box model. Future research can optimize the existing model by using more algorithms according to the national conditions of different countries, simplify the modeling process, and improve the prediction accuracy.

References

[1] Hong, J., Choi, H., & Kim, W. (2020). A house price valuation based on the random forest approach: The mass appraisal of residential property in South Korea. International Journal of Strategic Property Management, 24(3), 140–152.
[2] Time Is Gentle and Quiet. (2017). Analysis of factors affecting the price of second-hand housing in Beijing based on Lasso and data mining (Master's thesis). Beijing University of Technology. (In Chinese) (Note: Pseudonyms or unique online handles used as author names are capitalized as proper nouns per APA rules).
[3] Shi, A. H., Lu, L., & Yu, K. Y. (2026). Comparative study on real estate value prediction model based on random forest and deep neural network. Computer Age, (1), 47–51, 57.
[4] Hao, C. Y. (2020). Research on second hand housing transaction price prediction based on ensemble learning (Master's thesis). Shanxi University.
[5] Du, W. Y. (2024). Prediction of second-hand housing prices based on stacking fusion model and SHAP feature analysis (Master's thesis). Lanzhou University.
[6] De Cock, D. (2011, November 1). Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project. Kaggle. Retrieved March 16, 2026, from
[7] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
[8] Tao, Y., Xu, Z. C., & Tan, J. B. (2026). Prediction and driving mechanism analysis of soil organic carbon storage in Qinghai Tibet Plateau based on XGBoost-SHAP coupling model. Environmental Science, 1–17. Advance online publication.
[9] Yin, H., Alizamir, M., Heddam, S., et al. (2026). Improving prediction accuracy of mechanical strengths in high-performance concrete using advanced metaheuristic-based extreme learning machines: Incorporation of SHAP and LIME model explainers. Ain Shams Engineering Journal, 17(6), Article 104136.
[10] Wang, S. C., & Xu, Y. F. (2026). Research on the bias governance framework of inclusive finance algorithm based on the integration of SHAP and LIME. China Collective Economy, (12), 141–144.
[11] Song, J. (2022). Analysis of the impact of the quality of educational resources and school district policies on school district housing prices (Master's thesis). Southwestern University of Finance and Economics.

DOI: 10.54254/2753-8818/2026.DL34409

[12] Yang, Y. L. (2025). Research on the second hand housing price evaluation model based on SSA-XGBoost — Taking Xi'an Qujiang New Area as an example. China Collective Economy, (36), 85–88.
[13] Liang, Q. Y., Zhong, R., & Liu, B. (2026). Research on the spatio temporal differentiation characteristics and influencing factors of residential housing prices — Based on the multi-source data analysis of second-hand housing in Guangzhou. Shanghai Real Estate, (3), 56–62.