Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU34304

A Comparative Study of House Price Prediction Using Multiple Models

Yang Ping Chen

School of Computer Science and Technology, Soochow University, Suzhou, China
a0975600654@gmail.com

Abstract. Predicting housing prices, therefore, has important implications for real estate management, financial risk assessment, and urban planning. The rise of big data has further opened up new tools for housing price prediction, enabling the use of more complex and data-driven approaches alongside traditional methods. This study applies linear regression, decision tree regression, and random forest regression to the Boston housing dataset. Model performance is evaluated using MAE, MSE, and R^2. The random forest model achieves the best performance (MAE=2.05, MSE=7.93, R^2=0.89), outperforming linear regression (R^2=0.67) and decision tree regression (R^2=0.86).Correlation analysis reveals that the number of rooms (RM) shows the strongest positive correlation (r=0.70), while the lower-status population proportion (LSTAT) exhibits the strongest negative correlation (r=–0.74), together with crime rate (CRIM), these features form a core explanatory framework. The proportion of low-income population (LSTAT), and the distance to employment centers (DIS) are the most significant factors affecting housing prices. This study not only provides effective machine learning methods for housing price prediction but also offers data support for real estate appraisal, credit risk control, and urban policy making.

Keywords: housing price prediction, machine learning, feature importance analysis, regression model

1. Introduction

Housing price is an important indicator reflecting urban economic development and quality of life of residents. Research on housing price prediction carries significant value for real estate management, financial risk assessment, and urban planning. Traditional approaches to housing price prediction primarily rely on statistical regression methods. However, with the improvement of data availability and computational power, machine learning methods have gradually demonstrated stronger predictive performance and greater adaptability [1].

Much of the existing research has focused on improving single models, such as support vector machines and neural networks, for housing price prediction [2]. However, systematic comparisons across multiple models remain limited, and the interpretability of different features has not yet been sufficiently explored. The Boston housing dataset, as a classic empirical dataset, has been widely used in regression modeling, but there is still considerable room for deeper analysis of feature importance.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GU34304

This study uses the publicly available dataset and applies three machine learning models: linear regression, decision tree regression, and random forest regression. They are used to conduct a comparative analysis and identify the key factors that influence housing prices. By combining multi-model comparison with feature importance analysis, the paper aims to improve model interpretability and offer a feasible framework for housing price prediction. The findings hold practical value for real estate market analysis, financial risk control, and policy formulation.

2. Literature review

Earlier research on housing price prediction predominantly employed multiple linear regression models. With the advancement of ML, models have gradually been introduced into economic feild. Truong et al. conducted housing price prediction research using improved machine learning techniques and found that ensemble learning method significantly outperformed traditional regression models in terms of prediction accuracy [2]. Ping et al. analyzed factors affecting housing prices using multiple linear regression, validating the effectiveness of traditional regression methods for cases where feature variables show strong linear relationships [3].

Subsequent studies have employed random forest to predict housing prices in the United States, demonstrating its superiority over traditional regression models. Through feature importance analysis, these studies identified building structure and location as key variables affecting housing prices. Xiang further combined a hybrid kernel SVM with the MIDAS model to forecast the real estate prosperity index, highlighting the advantages of machine learning methods in handling time-series and nonlinear features [1]. Despite these advances, existing studies tend to focus on the predictive performance of individual models, and systematic comparisons across different model types on a common dataset remain relatively scarce. Moreover, while feature importance is often reported, the explanatory mechanisms behind how specific features drive housing price variations have not been thoroughly explored. In particular, the integration of multi-model comparison with interpretable feature analysis, beyond simply ranking variable importance, remains an area that warrants further investigation. As a result, current research still faces limitations in both the systematic comparison of models and the explanatory depth regarding feature influences.

3. Research methods

3.1. Data processing

3.1.1. Data selection

This study uses the Boston housing dataset, which contains 506 observations with 14 feature variables and a target variable MEDV (median home value in thousands of dollars). The main features includes indicators such as crime rate, number of rooms, proportion of low-income population, and pupil-teacher ratio, which reflect community socioeconomic conditions and building structure characteristics.

3.1.2. Data preprocessing

To ensure data quality and model robustness, a streamlined preprocessing pipeline was applied to the Boston housing dataset, primarily identifying and handling missing values, outliers, and duplicate records to prevent these issues from interfering with model training [4].

DOI: 10.54254/2755-2721/2026.GU34304

Regarding missing value handling, a completeness check of the dataset confirmed that there were no missing values. Outlier detection was performed using the boxplot method on key features highly correlated with housing prices, including the number of rooms (RM), the proportion of low-income population (LSTAT), and crime rate (CRIM). Identified extreme values were processed using the capping method. CHAS (whether the tract bounds the Charles River) is a binary variable and was used without encoding. Z-score standardization was applied to eliminate the dimensional effects among different features, ensuring that each feature contributes equally to model training. The preprocessed data were randomly split into training (80%) and test (20%) sets with a fixed random seed (42).

3.2. Model selection and setup

To systematically evaluate the performance of different modeling strategies in housing price prediction, this study selects three representative regression models for comparison. These three models respectively represent linear models, nonlinear single models, and ensemble models. Their combined use enables a comprehensive assessment of the performance of different modeling approaches in housing price prediction, providing a multi-dimensional reference for model selection.

3.2.1. Linear regression

Linear regression was selected as the baseline model. It predicts housing prices by fitting a linear relationship between features and the target variable, allowing for an intuitive interpretation of the direction and magnitude of each factor's influence [3]. It was implemented using ordinary least squares (OLS) with default parameters in Scikit-learn.

3.2.2. Decision tree regression

Decision tree regression is capable of automatically capturing nonlinear relationships and interactions [5]. The model recursively partitions the feature space and predicts the mean of the samples in each leaf node. It was implemented with Scikit-learn's default settings (random seed=42, min_samples_split=2, min_samples_leaf=1, criterion=MSE).

3.2.3. Random forest regression

Random forest regression, as an ensemble learning method, constructs multiple decision trees and aggregates their predictions. This approach effectively reduces the risk of overfitting which is common in single decision trees, while improving model generalization powe and prediction stability [6]. The model was implemented using Scikit-learn with 100 decision trees based on bootstrap sampling. All other parameters were set to default values.

These three models represent linear models, nonlinear single models, and ensemble models, respectively. Their joint application enables a comprehensive evaluation of the performance of different modeling approaches for housing price prediction, providing a multi-dimensional reference for housing price model selection.

3.3. Evaluation metrics

Model performance is evaluated using three standard metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and the coefficient of determination (R^2) [7].

DOI: 10.54254/2755-2721/2026.GU34304

Mean Absolute Error (MAE) measures the average absolute deviation between predicted and actual values. It is easy to interpret and less sensitive to outliers, directly reflecting the typical prediction error.Mean Squared Error (MSE) computes the average of squared errors, penalizing large deviations more heavily. It provides a sensitive measure of overall prediction accuracy.Coefficient of Determination (R^2) indicates the proportion of variance in the target variable explained by the model. Values closer to 1 imply better fit, showing how well the model captures the underlying patterns.

4. Results

4.1. Histogram distribution of MEDV (median housing price)

The distribution of housing prices (MEDV) in the Boston area is shown in Figure 1. The data indicate that housing prices are primarily concentrated from $20,000 to $25,000, with the distribution exhibiting a mild left-skewed shape. This suggests that most housing prices are at moderate levels, while a small number of high-end properties raise the upper limit of the overall price distribution. This distribution pattern is further supported by descriptive statistics: the average housing price is $22,530, the median is $21,200, and the standard deviation is $9,190.

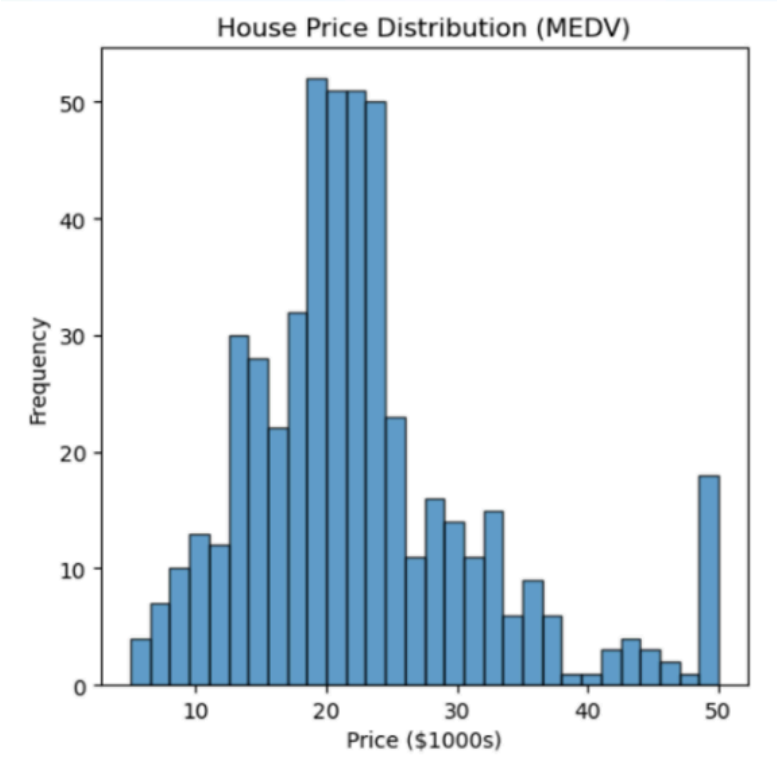


Figure 1. House price distribution (MEDV)

DOI: 10.54254/2755-2721/2026.GU34304

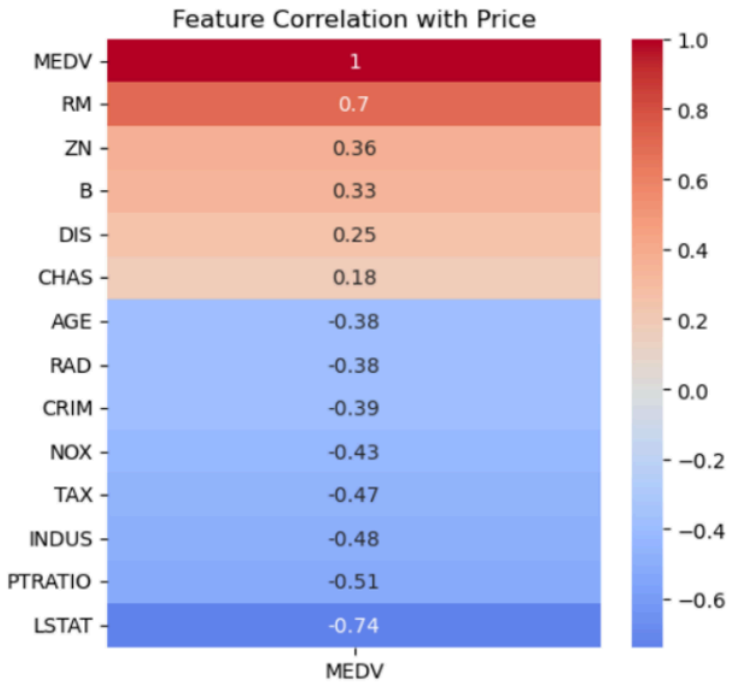


Figure 2. Correlation strength between various factors and boston housing prices

As shown in Figure 2, the number of rooms (RM) exhibits the strongest positive correlation with housing prices (a coefficient of approximately 0.70), while the proportion of low-income population (LSTAT) shows the most significant negative correlation (a coefficient of approximately -0.74). These three features, RM, CRIM, and LSTAT, collectively form a foundational framework for understanding and predicting Boston housing prices, representing the dimensions of building structure, community safety, and socioeconomic conditions. Among them, the positive effect of RM and negative effect of LSTAT are particularly prominent, making them the most critical explanatory variables in the modeling process.

In addition, environmental and cost factors such as crime rate (CRIM), nitrogen oxide concentration (NOX), and property tax rate (TAX) also exhibit stable moderate negative correlations with housing prices. The models clearly shows the key negative influencing dimensions of housing prices.

4.2. Model prediction visualization

Figure 3 visually compares the prediction performance of three models. The prediction points of the random forest model are most tightly clustered around the ideal diagonal line. Particularly in the mid-to-high price range, the deviation between predicted and actual values is significantly smaller than that of the other two models.

Linear regression exhibits systematic deviations in both high-price and low-price intervals, failing to capture nonlinear relationships. The decision tree prediction has a stepped distribution, indicating strong local fitting but insufficient overall smoothness. Random forest prediction points are tightly distributed around the 45° diagonal line, with notably lower deviations in the high-price region (MEDV> 30), confirming that its ensemble mechanism can effectively capture complex patterns.

DOI: 10.54254/2755-2721/2026.GU34304

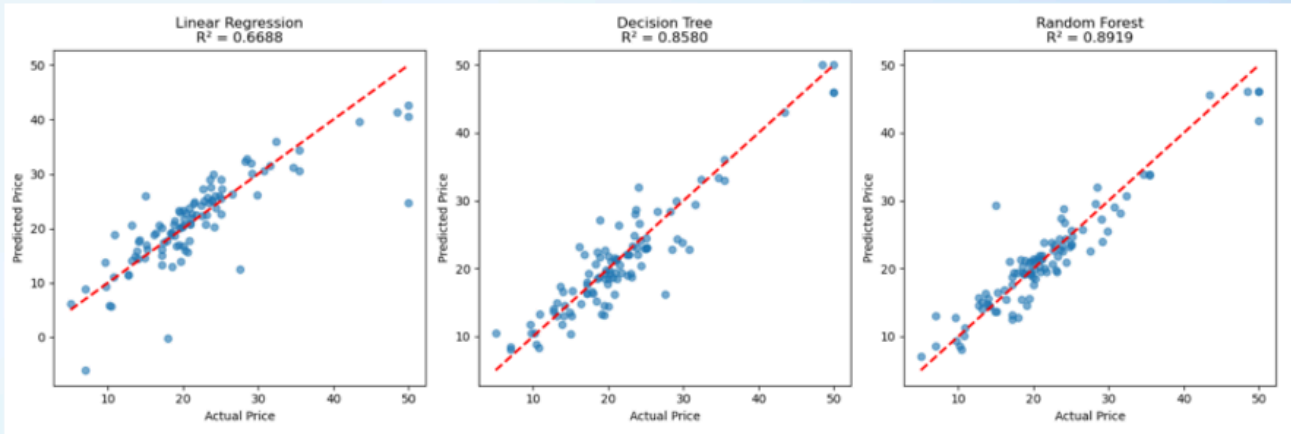


Figure 3. Scatter plot of predicted vs. actual values for three models

4.3. Model performance comparison

Table 3. Performance comparison of different models on the test set

Model	MAE	MSE	R^2
Linear Regression	3.1891	24.2911	0.6688
Decision Tree	2.3941	10.4161	0.8580
Random Forest	2.0474	7.9271	0.8919

The random forest model achieves the best performance across all evaluation metrics. It has the lowest mean absolute error (MAE = 2.05), indicating the smallest average deviation between predicted and actual values; the lowest mean squared error (MSE = 7.93), demonstrating higher prediction accuracy and substantially reducing the impact of large errors. The highest coefficient of determination (R^2 = 0.89), which implies that the model explains approximately 89% of the variance in housing prices and yields the best fit among the three models [9].

In contrast, linear regression shows significantly weaker predictive performance (MAE = 3.19, R^2 = 0.67), indicating its difficulty in capturing complex relationships within the data. While the decision tree outperforms linear regression, its metrics are inferior to those of random forest, reflecting the potential instability of single-tree models.

4.4. Model validation

To visually demonstrate the predictive capability of the optimal model (random forest), a single test sample was randomly selected for point prediction. The results show that for a property with an actual price of $23,600, the model predicts a value of $22,849, yielding a prediction error of $751 and a relative error of approximately 3.2%. This indicates that by leveraging the advantages of ensemble learning, the random forest model excels in accuracy, robustness, and interpretability.

DOI: 10.54254/2755-2721/2026.GU34304

5. Discussion

5.1. Findings

According to the results, random forest got the best overall predictive performance, with a mean absolute error (MAE) of 2.05, a mean squared error (MSE) of 7.93, and a coefficient of determination (R^2) of 0.89, significantly outperforming the other two models. Linear regression yields markedly weaker results (MAE = 3.19, R^2 = 0.67), indicating its inability to capture the nonlinear relationships present in the data. The decision tree model falls between the two, with its metrics (MAE = 2.39, R^2 = 0.86) reflecting the potential instability of single-tree structures. the correlation analysis reveals a clear structure of determinants. the random forest model achieves the best predictive performance overall, with mean absolute error (MAE = 2.05), mean squared error (MSE = 7.93), and coefficient of determination (R^2=0.89) significantly outperforming the other models. This indicates that, leveraging the advantages of ensemble learning, the random forest model leads in accuracy, robustness, and interpretability, making it the most reliable and closest-to-reality model for housing price prediction tasks [9]. Furthermore, the left-skewed distribution of housing prices in the Boston area indicates that the dataset is primarily composed of moderately priced properties, with a small number of high-end properties (right tail). This characteristic suggests the need to consider the potential impact of high-value samples on model parameter estimation, favoring the use of more robust algorithms [8]. These findings suggest that the model can be applied to assist in real-time real estate appraisal, mortgage loan risk assessment, and policy impact simulation across multiple scenarios.

5.2. Challenges

This study is limited by the small sample size of the Boston housing dataset, and future research should incorporate larger-scale time-series data. In addition, due to the lack of macroeconomic and policy variables in the dataset, future work could construct a multi-source data fusion framework. Since interactive features were not constructed in the experiments, future research could introduce advanced algorithms such as XGBoost and deep neural networks. The "black box" nature of random forests further limits its application in high-risk scenarios, and future efforts should attempt to introduce SHAP values to enhance model transparency [10].

Through multi-model comparison and feature interpretation, this study not only validates the superiority of ensemble learning methods in housing price prediction but also provides an analytical framework that combines prediction accuracy with interpretability for real estate appraisal, urban policy formulation, and financial risk control.

6. Conclusion

Using the Boston housing dataset, this study uses linear regression, decision tree regression, and random forest regression to predict housing prices and explore their key influencing factors. The results demonstrate that the random forest model achieves the best performance, with an R^2 of 0.89, MAE of 2.05, and MSE of 7.93, clearly outperforming both the linear regression (R^2 = 0.67) and the decision tree (R^2 = 0.86). Feature importance analysis further reveals that the number of rooms (RM) exhibits the strongest positive correlation (r = 0.70), while the lower-status population proportion (LSTAT) shows the strongest negative correlation (r = –0.74). These findings confirm the advantage of ensemble learning methods in housing price prediction and offer an interpretable

DOI: 10.54254/2755-2721/2026.GU34304

analytical framework that can support real estate appraisal, risk assessment, and policy-making. Feature importance analysis further reveals that the number of rooms and the proportion of LSTAT are the most critical positive and negative factors influencing housing prices. This study validates the superiority of ensemble learning methods in housing price prediction and provides an interpretable analytical framework for real estate appraisal.

Howevr, the dataset is constrained by its temporal and regional scope, which prevents it from fully reflecting contemporary market characteristics. Second, the relatively small sample size may affect the model's generalization ability. Future research can be extended in the following directions: developing advanced algorithms and collecting more recent housing price data; further refining feature engineering by constructing indicators such as location convenience; employing SHAP values to enhance model interpretability; and exploring spatial heterogeneity in housing prices to provide a basis for differentiated policy making.

References

[1] Xiang, C., & Hans. (2018). Research on real estate prosperity index prediction using SVM with mixed kernel function to correct the MIDAS model. Journal Name, 7 (12).
[2] Truong Q, Nguyen M, Dang H, Mei B. Housing price prediction via improved machine learning techniques. Procedia Computer Science, 174, 433–442.
[3] Ping, Y. L., & Hans. (2020). Analysis of the impact on construction prices based on multiple linear regression. Journal Name, 9(1).
[4] Wu, X., Dong, B., Du, X., & Yang, W. (2019). Data governance technology [Technical report].
[5] Liang, S., Chen, J., Wu, B., & Chen, G. (2015). Extracting rubber plantation information using object-oriented decision tree model. Journal of Remote Sensing, 19(3).
[6] Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. Journal Name, 18.
[7] Wang, J., Ran, Q., Deng, H., Huang, G., & Wang, X. (2018). Stand growth model of Pinus tabuliformis based on error-in-variable method. Journal of Zhejiang A&F University, 35(4).
[8] Zhang, J., & Guo, H. (2024). A review and outlook on data-driven mass appraisal of real estate. Journal of Xihua University (Philosophy & Social Sciences), 43.
[9] Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32.
[10] Wu, D. (2024). Application of big data technology in enterprise accounting data analysis. China Science & Technology Investment, (5), 27–29.