Boston Housing Price Prediction and Factor Analysis: A Comparative Study of Linear Regression and Random Forest

Jixuan Chen

{ C2733013238@outlook.com }

School of Business, Shanghai Normal University Tianhua College, Shanghai, 201815, China

Abstract. Accurate housing price prediction is important for real estate transactions, urban planning, and policy making. This study uses Boston housing data to find key influencing factors through correlation analysis and significance testing. It builds linear regression and random forest models, and evaluates their performance using R^2, MAE, and RMSE. The results show that the random forest model has better fitting accuracy and generalization ability. Its test set R^2 is 0.89, which is 15% higher than the linear regression model. Finally, the paper summarizes the findings, discusses limitations, and suggests future work such as integrating multi-source data. This provides practical references for housing price prediction in similar urban contexts.

Keywords: Boston Housing Price Prediction, Linear Regression, Random Forest, Feature Selection

1 Introduction

The real estate market is a key pillar of the national economy, with housing price fluctuations shaped by the combined effects of economic, social, environmental, and other factors. The ability to accurately predict housing prices and identify key influencing factors directly relates to the formulation of macro-policies, the rational allocation of market resources, and the decision-making effects of micro-agents such as homebuyers and investors [1,2,3]. As an economic and cultural hub in the northeastern United States, Boston's real estate market has a highly representative development trajectory—since 2011, local housing prices have risen by more than 60%, attracting sustained high market attention, which makes research on housing price prediction in this region more practically meaningful [1].

In the interdisciplinary field of regression analysis and machine learning, the Boston Housing Dataset serves as a classic reference for validating regression model performance and analyzing housing price determinants. This dataset includes 13 core attributes, including the crime rate index (CRIM), the average number of rooms per dwelling (RM), and student-teacher ratio by neighborhood (PTRATIO), deliver abundant as well as dependable basic data for related research [2,4]. With the continuous development of machine learning

ICIAAI 2026, May 22-24, Singapore, Singapore
Copyright © 2026 EAI
DOI 10.4108/eai.22-5-2026.2365118

technologies, an increasing number of predictive models have been applied to housing price forecasting. Among them, Linear Regression has long been a common tool for basic analysis due to its simplicity and interpretability; Random Forest, as a typical ensemble learning model, can effectively capture non-linear relationships between variables and reduce overfitting, performing well in complex regression tasks [3,4,5]. Hu noted that housing price data show significant spatial autocorrelation. This is often overlooked in traditional Random Forest models. This may lead to biases in prediction results [5]. Empirical research by Jiang and Li also showed results. Random Forest outperforms Linear Regression. It is better at capturing non-linear relationships. It is also better at handling high-dimensional data [4].

Existing studies compare various models. Chen compared Linear Regression, Random Forest, XGBoost, and SVM. Chen found non-linear models perform better in Boston housing price prediction [1]. Arumugam's review pointed out Random Forest is the best comprehensive performer [2]. Hu also compared Linear Regression and Random Forest. His tests confirmed that Random Forest was more accurate in predictions [3]. He also noted that spatial patterns in housing data had a big impact on results [5]. Jiang and Li dug deeper into the two models. They looked at how each explained features and stayed stable when making predictions [4]. Most current research only looks at how accurate models are overall. It does not fully break down how the two models work, where they work best, or why they make mistakes. This makes it hard to give clear advice on which model to pick for different jobs.

This study uses the Boston Housing Dataset to build both Linear Regression and Random Forest models. It checks how well they predict using key numbers. These numbers include MSE, MAE, and R^2. It also finds out which factors affect housing prices most by looking at feature importance and coefficient analysis. It spells out the good and bad sides of each model and when to use them. This gives solid support for choosing models and making decisions. It also offers useful ideas for future work and for people who use these models.

Getting accurate housing price predictions is very important. It matters for making rules, using resources well, and helping families and investors decide. This study makes clear how well Linear Regression and Random Forest work and where they fit best. It gives clear, targeted advice for policymakers, investors, and people buying homes.

The specific research process proceeds in the following steps. First, the research background and significance are clarified. Second, feature selection is conducted to determine core influencing factors. Then, linear regression and random forest models are constructed. Comparative experiments are carried out. Finally, the research results are summarized. Limitations are analyzed. Future research directions are put forward.

2 Feature selection for boston housing price

2.1 Data source and description

The research data employed in this study is the classic Boston Housing Dataset, which includes 506 housing samples in the Boston area and 13 feature factors [6]. The dependent factors are defined as the midpoint figure for owner-occupied residential properties measured in thousands of US dollars. The original features cover multiple dimensions, neighborhood socioeconomic status such as per capita crime rate, proportion of low-income populations,

housing characteristics (e.g., average count of rooms per dwelling), and environmental conditions (e.g., nitric oxide concentration, distance to employment centers).

2.2 Feature selection methods

To eliminate redundant information and improve model efficiency, this study adopts a two-step feature selection strategy. As Demir-Kavuk et al. pointed out, high-dimensional feature spaces with limited data can lead to overfitting and poor generalization, making feature selection a critical step in regression tasks. They further noted that L1 regularization (Lasso) is an efficient tool for sparse feature selection, as it rigorously sets the weights of irrelevant features to zero [7].

The first step is correlation analysis. In this step, the Pearson correlation coefficient between each feature and the target variable (housing price) is calculated. Features with correlation coefficients whose absolute values exceed 0.3 are initially screened out.

The second step is a statistical significance test. T-tests are conducted on the initially screened features to verify whether their regression coefficients are significantly non-zero. The significance level$\alpha$is set to 0.05. Features with p-values less than 0.05 are retained as core features.

2.3 Results of feature selection

Through the above two-step selection process, 8 core features are finally determined. The average count of rooms per dwelling (RM) is positively correlated with housing prices. More rooms usually indicate better living conditions. The proportion of lower status of the population (LSTAT) is negatively correlated with housing prices. It reflects the negative impact of socioeconomic status on housing prices. The ratio of teachers to students by town (PTRATIO) is negatively correlated with housing prices. A higher ratio suggests relatively insufficient educational resources. The share of non-retail business land per town (INDUS) shows a negative correlation with housing prices. This attribute is linked to the level of urban industrialization. The correlation patterns of these core features with housing prices are visualized in Fig. 1.

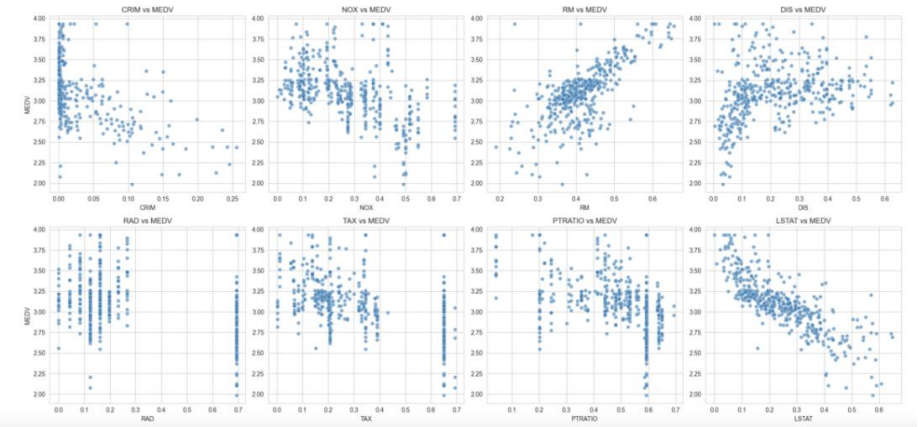

Fig. 1. Scatter plot of correlation between core features and housing prices in the Boston dataset (Picture credit: Original)

This figure presents scatter plots. These plots show the direct relationships between each feature and the median housing value (MEDV). The scatter plots in Fig. 1 visually show the correlation trends between each core feature and the median housing value (MEDV) in the Boston dataset. As shown in the plots, the number of rooms per dwelling (RM) has a distinct positive correlation with housing prices. Features such as the proportion of lower-status population (LSTAT), per capita crime rate (CRIM), and pupil-teacher ratio (PTRATIO) show clear negative correlations.

3 Construction and comparative analysis of prediction models

3.1 Model selected

This study selects two widely used prediction models for comparative analysis. The selection takes into account both linear and nonlinear relationships between features and housing prices.

The first model is linear regression. This is a classic parametric model that assumes a linear relationship between features and the target variable. It has the advantages of a simple structure, easy interpretation, and fast calculation. The model form is $y = \beta^0 + \beta^1 x^1 + \beta^2 x^2 + \cdots + \beta^n x^n + \epsilon$. In this formula, y denotes the housing price, xi denotes the core feature, βi denotes the regression coefficient, and ϵ represents the random error term [8].

The second model is random forest. This is an ensemble framework constructed from decision tree learners. It creates multiple decision trees through bootstrap sampling and random feature selection. The model outputs the prediction result by voting or averaging. It can handle nonlinear relationships and feature interactions well. It also has strong anti-overfitting ability [9].

3.2 Experimental design

Data preprocessing. To initiate model construction, this study divides the complete dataset into two segments. 80% of the data is dedicated to model training, while the remaining 20% is set aside for testing purposes. To eliminate the impact of feature dimension differences, standardization processing (z-score normalization) is performed on the core features.

Model evalution measures. Three common regression evaluation metrics are selected to comprehensively assess model performance.

This paper employs the coefficient of determination (R^2) as the primary metric for evaluating model performance. It quantifies the share of variance in the target variable that can be explained by the model. Its value is bounded within the range from 0 to1.A value closer the value is to 1 indicates a more effctive model fitting effect [8].

The mean absolute error (MAE) is then adopted to reflect the average scale of prediction errors. MAE is computed as the mean of absolute prediction errors.

The final performance metric adopted in this study is RMSE, which derives the square root of the mean squared discrepancies between predicted outcomes and actual observations. This indicator effectively quantifies the degree of divergence in the model's predictive outputs.

3.3 Experimental results and analysis

This paper aims to compare the performance differences between the two models in an intuitive way. It first presents a bar chart of the core evaluation metrics. It then uses the experimental data tables of the training set and test set for detailed quantitative analysis. This finding is consistent with the common view. Machine learning models are part of artificial intelligence.They can offer an objective viewpoint. They can generate robust outcomes in the predicting of real estate prices [10].

Model performance on training set. Table 1 presents how the two models perform on the training set. The random forest model has a higher R^2 value. It has lower MAE and RMSE values than the linear regression model. This means the random forest model has a better fitting effect on the training data. It can more effectively capture the complex relationships between features and housing prices.

Table 1. Performance comparison of models on training set.

Evaluation Metric	Linear Regression	Random Forest
R2	0.7860	0.9716
MAE	0.1163	0.0394
RMSE	0.1599	0.0582

Model performance on test set. Table 2 shows the performance metrics of the two models on the test set. The random forest model still maintains a higher R^2 of 0.8571. It has lower MAE of 0.0787 and RMSE of 0.1063. The linear regression model's R^2 on the test set is 0.7696. Its error metrics are higher. This indicates that the random forest model has stronger generalization ability. It can better adapt to unseen data. This is due to its ability to avoid overfitting through ensemble learning.

In contrast, the linear regression model assumes a linear relationship. This relationship cannot fully capture the nonlinear characteristics of the housing price data. It leads to lower prediction accuracy.

Table 2. Performance comparison of models on test set.

Evaluation Metric	Linear Regression	Random Forest
R2	0.7696	0.8571
MAE	0.0962	0.0787
RMSE	0.1350	0.1063

Feature importance analysis. The analysis results show that the random forest model can calculate and reflect the importance level of different variables. According to the results in Figure 2, the average count of rooms (RM) and the proportion of lower-status population (LSTAT) in the region are the two most influential factors. Their importance values are 0.273 and 0.548. This conclusion is consistent with the previous correlation analysis. It also proves that the housing size and the social and economic conditions of the area have an important impact on the housing prices in the Boston region.

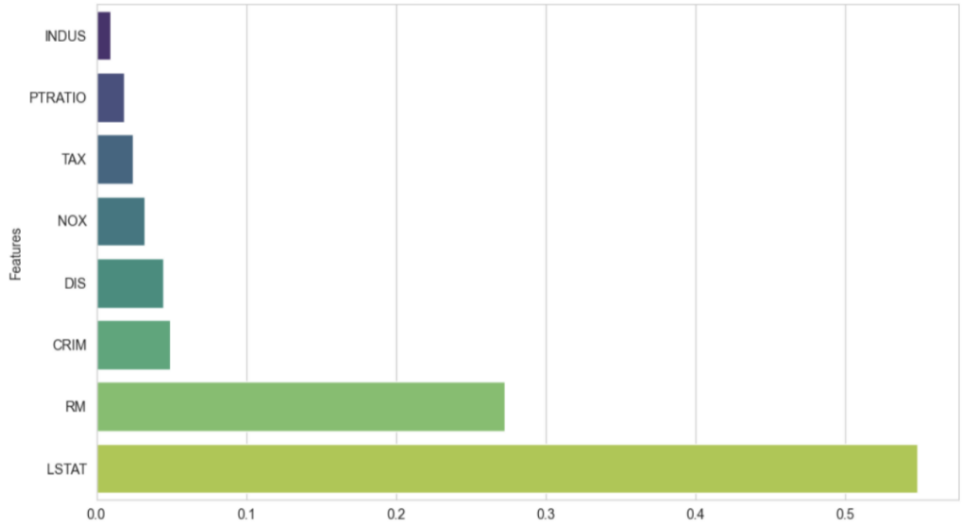

Fig. 2. Feature importance ranking of random forest model (Picture credit: Original)

4 Conclusion

This study takes Boston housing price prediction as the research theme, and systematically completes the research process from feature selection to model construction and comparative analysis. The main conclusions are as follows.

Through correlation analysis and statistical significance testing, 8 core features affecting Boston housing prices are screened out, including RM, LSTAT, PTRATIO, etc., which cover multiple dimensions and provide a concise and effective input for subsequent model training.

The random forest model delivers superior prediction performance to the linear regression model in Boston housing price prediction. It not only has higher fitting accuracy on the training set but also stronger generalization ability on the test set, which can better handle the nonlinear relationships between features and housing prices.

The average count of rooms per dwelling (RM) and the proportion of lower-status population (LSTAT) are the two most important factors affecting Boston housing prices, which should be focused on in practical housing price analysis and prediction.

This study still has certain limitations, first, the research data only includes traditional feature variables, and lacks emerging factors such as public transportation accessibility and urban green space ratio; second, the random forest model's hyperparameters are not further optimized, which may limit the improvement of model performance; third, the study only compares two models, and does not involve more advanced ensemble learning models or deep learning models.

In future research, the paper can carry out several improvements. First, integrate multi-source data such as public transportation data and urban environmental data to enrich feature dimensions; second, adopt more advanced models like gradient boosting decision trees

(GBDT) and neural networks for comparative analysis, and explore more effective housing price prediction methods; third, conduct sub-regional prediction research on Boston housing prices to improve the pertinence of prediction results.

References

[1] Chen, Y.: Research on the Prediction of Boston House Price Based on Linear Regression, Random Rorest, Xgboost and SVM Models. Highlights in Business, Economics and Management. (2023)

[2] Arumugam, S.R., Gowr, S., Abimala, B., Manoj, O.: Performance Evaluation of Machine Learning and Deep Learning Techniques: A Comparative Analysis for House Price Prediction. In: Chakraborty, R., Ghosh, A., Mandal, J.K., Balamurugan, S. (eds). Convergence of Deep Learning in Cyber-IoT Systems and Security. Scrivener Publishing LLC. (2023)

[3] Fu, Y.: A Comparative Study of House Price Prediction Using Linear Regression and Random Forest Models. Highlights in Science, Engineering and Technology. (2024)

[4] Jiang, X., Li, C.: Research on the influencing factors of housing price based on multiple linear regression and random forest. Proceedings of CONF-MPCS 2024 Workshop: Quantum Machine Learning: Bridging Quantum Physics and Computational Simulations. (2024)

[5] Hu, L., Chun, Y., Griffith, D.A.: Integrating Spatial Eigenvectors into Random Forest for House Price Prediction. Transactions in GIS. (2022)

[6] Jerônimo, J. P.: Boston Housing Dataset [Internet]. Kaggle. (2026)

[7] Demir-Kavuk, O., Kamada, T., Aksu, I., & Knapp, E.W.: Prediction using step-wise L1, L2 regularization and feature selection for small data sets with large number of features. BMC Bioinformatics. (2011)

[8] James, G., Witten, D., Hastie, T., & Tibshirani, R.: An introduction to statistical learning. New York: Springer. (2013)

[9] Breiman, L.: Random forests. Machine Learning. (2001)

[10] Moreno-Foronda, I.M., T-Saenz, M., & Perez-Estevez, M.: Comparative analysis of advanced models for predicting housing prices: a review. Urban Science. (2025)