Proceedings of CONF-BPS 2026 Symposium: Sustainability Transitions and Regional Economic Restructuring in New Contexts
DOI: 10.54254/2754-1169/2026.NJ30809

A Comparative Analysis of Machine Learning Algorithms for Forecasting Housing Prices in Boston

Yingxi Chen

Department of Finance, East China University of Science and Technology, Shanghai, China
24011929@mail.ecust.edu.cn

Abstract. Real estate is a crucial and multifaceted field that encompasses property valuation, investment, urban development, and even aspects related to the surrounding regional economy. Accurate predictions of housing prices in Boston are therefore of great significance to its various stakeholders. To address this issue, in this study, after thorough data preprocessing and detailed feature engineering, multiple machine learning models were employed to predict the housing prices in Boston. These models include linear regression (LR), random forest (RF), and extreme gradient boosting (XGBoost). These models cover linear algorithms, ensemble methods, and tree-based techniques, and can comprehensively assess their ability to simulate the relationship between the features and prices. Among these models, XGBoost performed the best, achieving the highest R-squared value (0.920), the lowest root mean square error (RMSE = 2.7831), and excellent cross-validation performance. This indicates that XGBoost is highly effective in predicting housing prices in Boston and can provide feasible and practical prediction solutions for stakeholders.

Keywords: Boston Housing, Price Prediction, Machine Learning, XGBoost, Regression Models

1. Introduction

The real estate market in Boston has always been a key research subject in real estate studies and policy analysis. This is not only because Boston is the educational, technological, and financial center of the northeastern United States, but also because of the challenges it has faced over the long term. These factors make it a standard case for studying housing prices. For example, in 1991, the severe collapse of the Japanese real estate market led to a significant drop in housing prices, triggering serious social unrest, which demonstrates the importance of studying housing prices. At the same time, some researchers have found that due to the continuous increase in the cost of living in Boston, people are increasingly inclined to choose alternative options that can save expenses in the long term, that is, directly purchasing a house [1].

Based on this dataset, many scholars have conducted research on it. Sharmila found that the neural network (NN) performed the best in controlling the error of housing price predictions, with its Mean Squared Deviation (MSD) (0.02776587) and Mean Absolute Deviation (MAD) (0.1274102) being much lower than those of decision trees (DT) and multiple linear regression (MLR) [2]. The Sinha used the root mean square error (RMSE) and other indicators to evaluate the

DOI: 10.54254/2754-1169/2026.NJ30809

performance of the prediction models, and ultimately found that the XGB model achieved excellent results with RMSE = 2.45, Mean Squared Error (MSE = 6.03), Mean Absolute Error (MAE=1.32), and Mean = 21.04, being the best performer [3]. In "Research on the Prediction of Boston House Price Based on linear regression (LR), random forest (RF), extreme gradient boosting (XGBoost), and support vector machine (SVM) Models", the author also found that the R-square score of the XGBoost regression model was 85.799520. It performed best among RF, LR, and support vector machines, being the most suitable model for this dataset [4]. Therefore, this paper will reprocess the data and use the model for calculation to observe which model is more suitable for this prediction problem.

2. Methodology

2.1. Data description

The Boston Housing Dataset contains information related to housing from 506 samples, which were collected in the 1970s and cover different areas of Boston [5]. This dataset includes 14 features: 13 numerical features (such as the crime rate per capita, the average number of rooms per house, the distance to the five employment centers in Boston, the property tax rate, the student-to-teacher ratio, and the accessibility index to the radioactive road) and 1 categorical feature (the Charles River dummy variable). The median value of the sold houses is the target variable of the study (in $1000s).

2.2. Data processing

This framework consists of several stages: data collection, preprocessing, training, testing, and algorithm implementation. The expected result lies in the final stage.

Before using the data for learning, it should be prepared by cleaning the data. Firstly, histograms and scatter plots are used to visualize the original data. This enables an intuitive understanding of individual data features and their relationship with house prices, which helps select subsequent data processing and modeling methods. Table 1 provides specific explanations for each indicator of the dataset used in the research. Then, statistical analysis of the data is conducted, as shown in Table 2. It can be seen that the distributions of INDUS, NOX, etc., are relatively concentrated, while CHAS, B, etc., have extreme outliers and require subsequent processing.

Table 1. Boston housing data: variable definitions

Abbreviation	Full Name & Meaning
CRIM	Per Capita Crime Rate by Town (per capita crime rate in the town)
ZN	Proportion of Residential Land Zoned for Lots Over 25,000 sq.ft. (share of large residential lots)
INDUS	Proportion of Non-Retail Business Acres per Town (share of non-retail business land)
CHAS	Charles River Dummy Variable (1 if tract bounds the river; 0 otherwise)
NOX	Nitric Oxides Concentration (parts per 10 million)
RM	Average Number of Rooms per Dwelling (average number of rooms in residential units)
AGE	Proportion of Owner-Occupied Units Built Prior to 1940 (share of old owner-occupied homes)
DIS	Weighted Distances to Five Boston Employment Centres (distance to employment hubs)

DOI: 10.54254/2754-1169/2026.NJ30809

Table 1. (continued)

RAD	Index of Accessibility to Radial Highways (highway accessibility index)
TAX	Full-Value Property-Tax Rate per $10,000 (property tax rate per $10,000)
PTRATIO	Pupil-Teacher Ratio by Town (student-teacher ratio in the town)
B	1000(Bk - 0.63)^2, where Bk is the Proportion of Black People by Town (derived measure of Black population share)
LSTAT	Percentage of Lower Status of the Population (share of low-income population)
MEDV	Median Value of Owner-Occupied Homes in $1000s (median home value, in thousands of dollars)

Table 2. Data description

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE
count	486	486	486	486	506	506	486
mean	3.612	11.212	11.084	0.070	0.555	6.285	68.519
std	8.720	23.389	6.836	0.255	0.116	0.703	28.000
min	0.006	0.000	0.460	0.000	0.385	3.561	2.900
25%	0.082	0.000	5.190	0.000	0.449	5.886	45.175
50%	0.254	0.000	9.690	0.000	0.538	6.209	76.800
75%	3.560	12.500	18.100	0.000	0.624	6.624	93.975
max	88.976	100.000	27.740	1.000	0.871	8.780	100.000
	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
count	506	506	506	506	506	486	506
mean	3.795	9.549	408.237	18.456	356.674	12.715	22.533
std	2.106	8.707	168.537	2.165	91.295	7.156	9.197
min	1.130	1.000	187.000	12.600	0.320	1.730	5.000
25%	2.100	4.000	279.000	17.400	375.378	7.125	17.025
50%	3.207	5.000	330.000	19.050	391.440	11.430	21.200
75%	5.188	24.000	666.000	20.200	396.225	16.955	25.000
max	12.127	24.000	711.000	22.000	396.900	37.970	50.000

Subsequently, box plots were employed to visually reveal the presence of outliers in certain features. As clearly observable in Figure 1, the “CRIM” variable (crime rate per capita) exhibits multiple outlying points significantly higher than the majority of data points, while the “B” variable (proportion of black residents in urban areas) also displays a small number of extreme values deviating from the main distribution. These outliers can distort statistical properties of the data—such as inflating the mean and widening the standard deviation—thereby obscuring the true distribution of features. More critically, they compromise the fitting performance of subsequent models, ultimately diminishing predictive accuracy.

The paper, therefore, addressed the outliers revealed by the box plot: employing the core logic for identifying outliers in box plots (the IQR rule, whereby values exceeding Q1-1.5×IQR or Q3+1.5×IQR are deemed outliers), we applied truncation to the “B” and “CRIME” indicators, confining their outliers within a reasonable range. Subsequently, we visually compared the distributions before and after processing using box plots to verify whether outliers were effectively

DOI: 10.54254/2754-1169/2026.NJ30809

controlled. This provided more reliable foundational data for subsequent analysis and modelling (Figure 2).

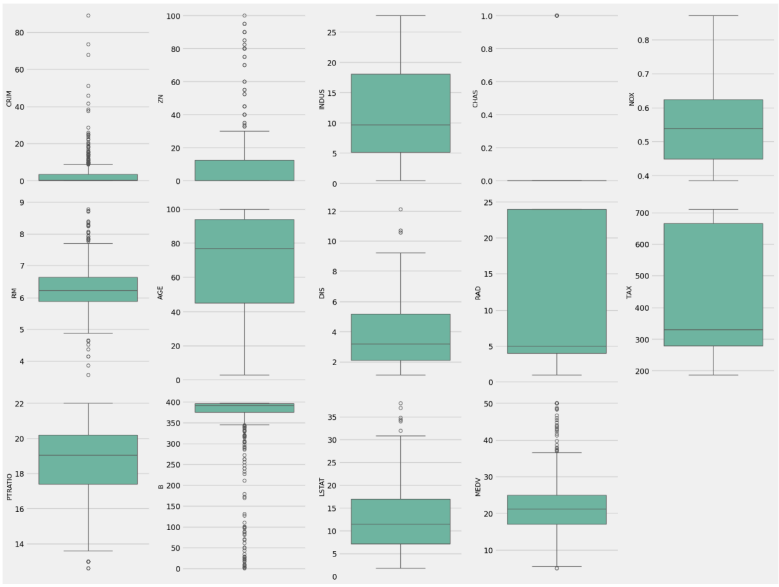

Figure 1. Box plots of Boston housing features (photo/picture credit: original)

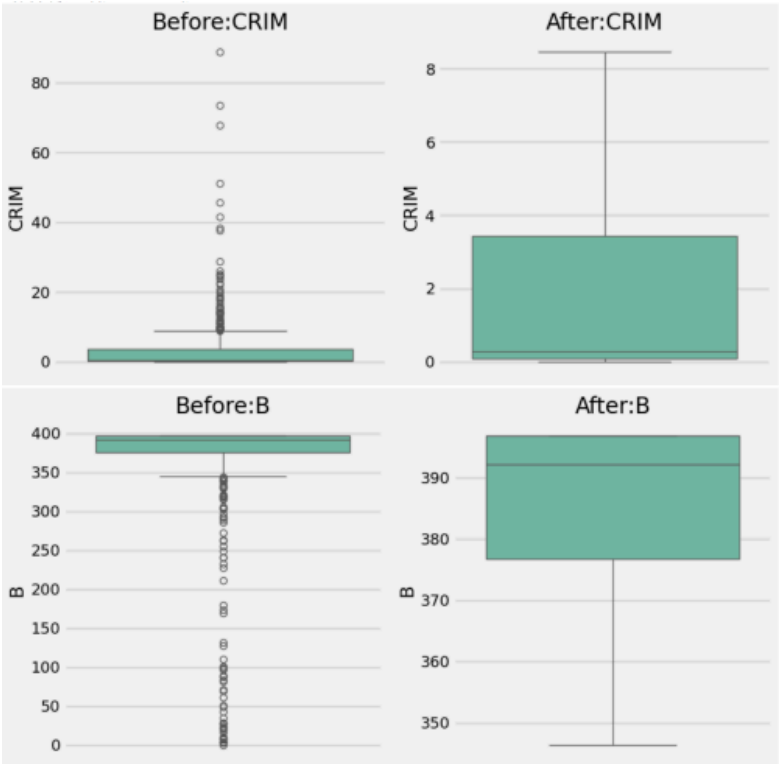

Figure 2. Box plots comparison of CRIM and B before and after outlier handling (photo/picture credit: original)

DOI: 10.54254/2754-1169/2026.NJ30809

2.3. Creative operation

Following the aforementioned foundational data preprocessing, the paper opted to employ multicollinearity analysis. After removing missing values from the dataset, the paper separated the feature matrix X (excluding the target variable MEDV) and the target variable y. Subsequently, the feature variables are standardised to ensure comparability among different dimensions. Finally, by calculating the variance inflation factor for each feature, the multiple linear correlation among the independent variables can be well evaluated. The VIF serves to measure the degree of multicollinearity among features, with higher VIF values indicating stronger multicollinearity.

For Table 3, the calculated VIF results indicate that the features TAX, RAD, and CRIM exhibit relatively high VIF values, signifying significant multicollinearity. To prevent instability in regression model coefficients and diminished interpretability, the paper has opted to remove these three high-VIF features using the drop method. This yields a more concise feature set, providing a more stable and interpretable data foundation for subsequent modelling.

Table 3. Vif

CRIM	ZN	INDUS	CHAS	NOX	RM	LSTAT
10.222032	2.359249	4.081733	1.071091	4.499886	2.068658	3.307986
AGE	DIS	RAD	TAX	PTRATIO	B	
3.165343	3.880238	13.137849	8.690021	1.847409	1.337022	

Next, the paper plots a heatmap like Figure 3 to visually assess the linear correlation between features and the target variable (property price MEDV), thereby providing a basis for subsequent feature selection, multicollinearity analysis, and model construction.

Figure 3. Correlation heatmap of Boston housing dataset features (photo/picture credit: original)

2.4. Algorithm

In the era of artificial intelligence, machine learning plays a significant role in predicting real estate prices. This paper addresses the task of predicting real estate prices (a regression problem) by extracting and identifying meaningful relationships from the housing dataset using LR, RF, and XGBoost.

DOI: 10.54254/2754-1169/2026.NJ30809

3. Results

3.1. Data stratification

Before building the prediction model, it is necessary to clean and preprocess the Boston housing price dataset and divide it according to the research needs. This study introduces two rounds of data splitting to boost the reliability and generalization performance of the model. In the first round of splitting, the original data is randomly divided into two parts: around 90% of the data is used as the main part for model development, and the remaining 10% is kept separately for the final model verification. Subsequently, further splitting is conducted on the model development part, with approximately 70% of it serving as the training set and the remaining 30% serving as the validation set, relied upon for performance comparison during model tuning.

3.2. LR model

First, the LR model was trained. Subsequently, predictions were made for the scaled features of both the intermediate validation set (comprising 27% of the total data) and the final validation set (comprising 10% of the total data). Metrics such as R^2, RMSE, and MAE were calculated to evaluate the model's performance on these two validation sets, with the assessment results outputted.

Next, to visually illustrate the prediction results, the paper plotted a scatter plot of actual vs. predicted property prices on the final validation set, adding a reference line (representing the ideal scenario where predicted and actual values perfectly align). Finally, the paper constructed a DataFrame containing actual property prices, predicted property prices, and corresponding errors, exporting all sample results to an Excel file.

$$\widehat{Y}=b_0+b_1X_1+b_2X_2+\ldots+b_kX_k \tag{1}$$

Among them, $(X_1, X_2, \ldots, X_k)$ are multiple independent variables, and $(b_1, b_2, \ldots, b_k)$ are the partial regression coefficients of each independent variable.

Figure 4 shows the comparison between the predicted house prices and the actual house prices by the linear regression model on the validation set. The horizontal axis represents the actual price of the house, and the vertical axis represents the predicted price output by the model. The red dotted line is the ideal fitting line where "predicted value = actual value". From the distribution characteristics, most of the scattered points corresponding to the predicted values are distributed around the baseline, indicating that the linear regression model can basically predict the overall trend of house prices; however, some scattered points are significantly deviated from the baseline, indicating that the fitting effect of this model is still insufficient. Figure 4 visually presents the prediction effect of the model and visualizes the prediction effect.

DOI: 10.54254/2754-1169/2026.NJ30809

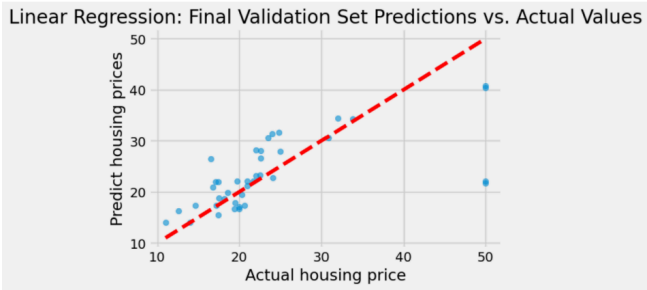

Figure 4. Results of the Linear regression model (photo/picture credit: original)

3.3. RF model and XGBoost model

Following the LR model, the paper proceeded to build RF and XGB models using the processed dataset and visualizing the results. See Figure 5 and Figure 6.

$$\mathrm{RF}: F(X) = \frac{1}{B}\sum_{i=1}^{B} T_i(X) \tag{2}$$

Where B is the number of trees, and Ti(X) is the predicted value of the i-th tree.

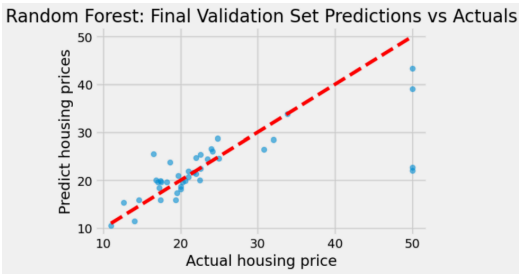

Figure 5. Results of the random forest model (photo/picture credit: original)

$$XGB: Obj^{(t)} = \sum_{j=1}^{T}\left(G_j\omega_j + \frac{1}{2}\left(H_j + \lambda\right)\omega_j^2\right) + \gamma T \tag{3}$$

Among them , $\left(G_j = \sum_{i\in I_j} g_i\right)$, $\left(H_j = \sum_{i\in I_j} h_i\right)$ (where (I_j) is the sample set of the j-th leaf node).

ωj² is the regularization term, T is the number of leaf nodes, ωj is the weight of the j-th leaf node, and γ and λ are regularization parameters

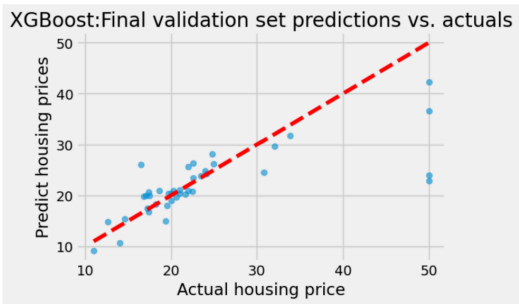

Figure 6. Results of the XGBoost model (photo/picture credit: original)

DOI: 10.54254/2754-1169/2026.NJ30809

3.4. Model Performance

Table 4. Performance of each model

	LR model	RF model	XGBoost model
R^2	0.7886	0.8913	0.902
RMSE	4.0868	2.9313	2.7831
MAE	2.9987	2.1406	1.9747

The performance of the LR, RF, and XGBoost models is well illustrated in Table 4 for the housing price prediction task. The model effectiveness is quantified by three indicators: R^2, RMSE, and MAE.

Among them, R^2 measures the model's ability to explain the data variation, while RMSE and MAE reflect the level of prediction error.

The results show that the XGBoost model is the best among the three for R^2($0.902 > 0.8913 > 0.7886$), and both RMSE (2.7831) and MAE (1.9747) are the smallest. This comparison clearly indicates that among the three models, XGBoost is more suitable for the complex correlation between housing prices and features, and has better prediction performance (Table 5).

Table 5. Comparison of predicted results with actual results

	LR model	RF model	XGBoost model
Error (abs'mean)	2.486	2.07	1.876
Error (variance)	4.859384	2.62274	1.103804

The paper used 90% of the data to build prediction models, then compared them against the remaining 10% of actual data to determine the deviation. Among the three models, XGB still exhibited the smallest deviation. This once again demonstrates that the XGB model performs best in the task of predicting housing prices.

3.5. Discussion

In this study, three machine learning models were employed to predict the house prices in Boston: LR, RF, and XGBoost. The XGBoost model performed the best, with a high R-squared value and a very low root mean square error.

However, this study also has some limitations: historical data may be biased and cannot reflect the recent market fluctuations, and the verification under real-time market conditions is also lacking. As Chiu, S. M mentioned in the article, Future research should focus on improving feature engineering, incorporating time and space-related factors, and promoting the development of model integration technology [6, 7]. Thus, it is essential to ensure that the research results apply to the current market conditions and can be extended to other cities, conducting practical tests using the latest data. Automated machine learning (AutoML) can simplify model exploration and hyperparameter adjustment, improve prediction accuracy, and reduce manual operations [8]. By adopting these advanced technologies for real estate price prediction, relevant parties can achieve more accurate valuations and provide data-based solutions for the Boston housing market and other similar urban environments.

DOI: 10.54254/2754-1169/2026.NJ30809

4. Conclusion

This study mainly employed three algorithms: LR, RF, and XGBoost. Among these algorithms, XGBoost performed the best, not only achieving the highest R-squared value, but also achieving the smallest MAE and RMSE, while successfully capturing the non-linear relationship between features and house prices. The cross-validation results indicated a strong consistency between the training results and generalization ability, which is crucial for reliable house price predictions in a dynamic real estate market. To address the research gap in the literature regarding the unique housing market characteristics of Boston (such as historical neighborhood patterns and urban development trends), this study comprehensively compared regression models and analyzed the importance of features in the local context. In addition to providing a new perspective for researchers in other regions applying machine learning to real estate datasets, the rigorous methodology in model selection and evaluation also improved the overall quality of urban real estate prediction research.

This study provides a reference for other researchers in other regions applying machine learning to real estate datasets, providing empirical evidence for the effectiveness of XGBoost in predicting house prices in Boston. Moreover, it offers practical insights for real estate professionals and policymakers to optimize pricing strategies, property assessment methods, and urban development plans. It has strong generalization ability and is highly consistent with the trained model. Compared with other models, its prediction performance is superior.

References

[1] Ding, H. (2024). Predicting Boston housing prices using machine learning models. Proceedings of the International Conference on Artificial Intelligence and Data Science, 439–444.
[2] Analysis and prediction of real estate prices: A case of the Boston housing market. (2018). Issues in Information Systems, 19(4), 12–19.
[3] Sinha, H. (2024). Benchmarking predictive performance of machine learning approaches for accurate prediction of Boston house prices: An in-depth analysis. International Journal of Advanced Research and Interdisciplinary Scientific Endeavours, 1(2), 78–91.
[4] Chen, Y. (2023). Research on the prediction of Boston house prices based on linear regression, random forest, XGBoost and SVM models. Highlights in Business, Economics and Management, 21, 27–37.
[5] Vishal, V. (2025, February 30). Boston housing dataset. Kaggle. https: //www.kaggle.com/datasets/altavish/boston-housing-dataset
[6] Chiu, S. M., Chen, Y. C., & Lee, C. (2022). Estate price prediction system based on temporal and spatial features and lightweight deep-learning model. Applied Intelligence, 52(1), 808–834.
[7] Boosting the accuracy of property valuation with ensemble learning and explainable artificial intelligence: The case of Hong Kong. (2023). The Annals of Regional Science.
[8] Tejaswi, V. S. B., & Satyanarayana, K. V. V. (2020). House sale price prediction using advanced regression techniques and AutoML (TPOTRegressor). International Journal of Emerging Trends in Engineering Research, 8(4), 1373–1378.