Available Online: 14 October 2025 DOI: 10.54254/3029-0880/2025.27554

Prediction of the Median Value of Owner-Occupied Homes (MEDV) in the Boston housing dataset using regression models

Xiaojin Zhou

School of Mathematics and Statistics, Shanghai Lixin University of Accounting and Finance, Shanghai, China

3462166338@qq.com

Abstract. The Boston housing dataset is one of the significant tools used to examine the influencing factors of housing price. Meanwhile, housing price prediction is crucial for government regulation, business decision-making, and individual homebuyers. Existing studies fall short in balancing model interpretability, computational efficiency, and generalization ability. Hence, this study, based on the Boston housing dataset, focuses on the prediction of the Median Value of Owner-Occupied Homes (MEDV). It constructs three models, including linear regression, decision tree regression, and Bayesian regression, and evaluates their performance using Mean Squared Error (MSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). It examines the effects 13 features, including Per Capita Crime Rate by Town (CRIM) and Average Number of Rooms Per Dwelling (RM) on MEDV. Furthermore, it trains and visualizes the results of each model. The results show that decision tree regression achieves the highest R^2, effectively capturing nonlinear relationships but being prone to overfitting. Linear regression and Bayesian regression perform better in terms of MSE and MAE; the former is simple in structure and fast to train, while the latter can output probability distributions to assess uncertainty. Each model has its strengths, and the choice should depend on the application scenario. The limitations related to dataset timeliness and the lack of extensive hyperparameter tuning are acknowledged, providing useful insights for housing price prediction research and practice.

Keywords: Boston housing dataset, regression model, MEDV prediction, model evaluation, feature engineering

1. Introduction

Since it was proposed by Harrison and Rubinfeld in 1978, the Boston housing dataset has become an important resource for studying the factors that influence housing prices [1]. As a core link between people's livelihoods and economic activity, accurate housing price prediction is crucial not only for government policies such as purchase restrictions and property tax adjustments, but also for real estate companies optimizing site selection, housing design, and pricing strategies. It can also help individual homebuyers mitigate the risks of price fluctuations and identify cost-effective properties, while providing feedback for urban planning. Analyzing features such as the Proportion of Lower-income Residents (LSTAT) and Pupil-teacher Ratio in Towns (PTRATIO) in relation to housing prices can support decisions on educational facilities and transportation planning, ultimately promoting the development of livable cities.

The Boston housing dataset, with its clear features and well-structured format, is widely used in both linear and nonlinear regression studies. Zhang et al. compared linear regression, ridge regression, and Support Vector Regression (SVR), finding that nonlinear models outperform traditional linear methods in handling complex feature relationships and controlling errors [2]. Wang et al. used random forest regression and identified RM and LSTAT as the most critical variables affecting the Median Value of Owner-Occupied Homes (MEDV) [3]. Chen et al. applied a Multilayer Perceptron (MLP) for prediction, showing that optimizing feature engineering can significantly enhance the neural network's ability to capture nonlinear patterns [4, 5].

In terms of feature engineering, Li et al. improved the performance of linear models through standardization, missing value handling, and polynomial feature construction [6]. Zhao et al. applied Principal Component Analysis (PCA) for dimensionality reduction to address multicollinearity, enhancing the generalization ability of certain regression models [7]. Despite extensive research, practical applications still face challenges in balancing model interpretability, computational efficiency, and generalization. Building on previous work, this study systematically reviews the performance of different regression models and, combined with feature engineering, further explores more effective approaches for predicting MEDV, providing targeted theoretical support and practical tools for housing price assessment, urban planning, and investment decision-making.

2. Methodology

2.1. Data sources and processing

2.1.1. Data sources

Proposed by Harrison and Rubinfeld in 1978, the Boston housing dataset includes 13 feature variables that affect housing prices and one target variable for prediction. The main features involve Per Capita Crime Rate by Town (CRIM), Proportion of Residential Land Zoned for Large Lots (ZN), Proportion of Non-retail Business Acres Per Town (CHAS), proximity to the Charles River, Nitric Oxide Concentration (NOX), Average Number of Rooms Per Dwelling (RM), Proportion of Owner-occupied Units Built Prior to 1940 (AGE), Weighted Distances to Employment Centers (DIS), Accessibility to Radial Highways (RAD), Full-value Property Tax Rate per $10,000 (TAX), Pupil-teacher Ratio by town (PTRATIO), population proportion by race (B), and Proportion of Lower-income Residents (LSTAT).

The target variable, MEDV, represents the median value of owner-occupied homes in thousands of dollars and is the primary focus of prediction in this study. The dataset has been standardized, which eliminates the effects of differing scales among feature variables and helps improve the training efficiency and convergence speed of subsequent regression models. In addition, the dataset contains very few missing values, so complex imputation is not required. The data quality is high overall, meeting the requirements for regression modeling and providing a solid foundation for the subsequent construction of accurate MEDV prediction models.

2.1.2. Data preprocessing

This study designs a systematic data preprocessing workflow tailored to the feature characteristics of the Boston housing dataset and the requirements of regression modeling. The workflow optimizes data quality through three core steps: standardization, dataset splitting, and feature selection. The specific procedures and underlying principles are as follows.

The 13 feature variables in the Boston housing dataset exhibit significant differences in scale. For example, CRIM typically ranges from 0.01 to 100, RM is concentrated between 3 and 9, and TAX can reach 100-700. Such scale inconsistencies can cause the model during training to be overly sensitive to features with large numerical ranges while neglecting those with smaller ranges, which in turn affects parameter update efficiency and model convergence speed.

Therefore, all feature variables in this study are standardized using Z-score normalization. This transforms each feature to a unified scale with a mean of 0 and a standard deviation of 1, eliminating the interference caused by differences in scale. It ensures that the model learns more balanced weights for all features, accelerates parameter convergence, especially for gradient descent-based linear regression models, prevents parameter oscillations caused by large differences in feature values, and improves overall training efficiency.

For feature selection, this study employs Pearson correlation analysis to evaluate each feature's relationship with MEDV. Based on the correlation coefficients, features with an absolute correlation below a set threshold, such as 0.3, are removed, while highly correlated features, such as RM, LSTAT, MEDV, and CRIM, are retained.

To evaluate the generalization ability of the models, the preprocessed Boston housing dataset is randomly split into training and test sets at a 70:30 ratio. This helps prevent overfitting, where the model might otherwise learn noise or specific patterns from the training set. During the splitting process, a fixed random seed is used to ensure reproducibility, while maintaining the feature distributions and MEDV distribution in the training and test sets consistent with the original dataset. This prevents distortion of model evaluation results due to biased data splitting and ensures the objectivity of subsequent model performance comparisons.

2.2. Research methods

2.2.1. Linear regression

Linear regression assumes a linear relationship between the target variable and the features. It is highly interpretable, computationally efficient, and easy to implement, making it a common baseline model for housing price prediction. The model can be expressed as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon \tag{1}$$

In this equation, y represents the target variable (MEDV), x_i represents the feature variables (such as CRIM, RM, LSTAT, etc.), β_i represents the regression coefficients, and ϵ is the error term. Linear regression estimates the model parameters using the least squares method and is suitable for data where a linear relationship exists between the features and the target.

2.2.2. Decision tree regression

Decision tree regression recursively partitions the feature space, dividing the data into multiple subsets and predicting a constant value (such as the mean) within each subset. Its advantage lies in capturing nonlinear relationships and interactions among features, making it suitable for complex datasets. However, this method is susceptible to overfitting and requires optimization techniques such as pruning.

2.2.3. Bayesian regression

Bayesian regression treats the regression coefficients as random variables with prior distributions and estimates model parameters by updating their posterior distributions through Bayesian inference. This method can provide information on prediction uncertainty and is suitable for scenarios with small sample sizes or where probabilistic interpretation is needed. Common Bayesian regression models include Bayesian ridge regression and Automatic Relevance Determination (ARD) regression.

3. Results

3.1. Performance evaluation of different regression models

To comprehensively assess the performance of the three regression models in predicting MEDV, this study uses three metrics, MSE, MAE, and R^2, to evaluate, from different perspectives, the deviation of the predicted values from the true values and the overall goodness of fit (Table 1).

Table 1. Comparison of performance evaluation metrics of three regression models

Model	MSE	MAE	R^2
Linear Regression	24.3691	3.2136	0.6677
Decision Tree Regression	10.4161	2.3941	0.8580
Bayesian Regression	24.4269	3.1774	0.6669

3.2. Analysis of model applicability

3.2.1. Applicability of linear regression models

The linear regression model offers strong interpretability and computational efficiency, making it suitable for scenarios where features have a clear linear relationship with MEDV, or when a quick baseline prediction or clear understanding of feature effects is needed. In the Boston dataset, RM is positively correlated with MEDV, while CRIM and LSTAT are negatively correlated, consistent with the linearity assumption. The model's regression coefficients provide an intuitive measure of influence—for example, a positive and relatively large coefficient for RM indicates that for each additional room, MEDV increases by several thousand dollars, offering clear guidance for government regulation, real estate pricing, and homebuyer decision-making. In addition, the data have been standardized using Z-score normalization to eliminate scale differences (for example, CRIM ranges from 0.01-100, TAX from 100-700), meeting the linear regression requirement for consistent feature scales. With a moderate number of features, namely 13, the model parameters can be efficiently estimated using the least squares method, resulting in low computational cost, making linear regression suitable as a baseline model to quickly validate linear trends.

However, linear regression assumes a linear relationship between the features and MEDV, making it difficult to capture nonlinear patterns and interaction effects. For example, Nitric Oxide Concentration (NOX) exhibits a threshold effect on housing prices: its impact is weak when the concentration is below 0.5, but prices decline rapidly as pollution exceeds 0.5, something a linear model cannot capture. Similarly, the interaction between RM and AGE (house age) matters: for new houses, more rooms significantly increase prices, whereas for older houses, even with more rooms, the price increase is limited due to outdated facilities. A linear model cannot recognize this synergistic effect, potentially leading to overestimation of older houses and reducing overall prediction accuracy.

3.2.2. Applicability of decision tree regression models

Decision tree regression recursively partitions the feature space, effectively capturing nonlinear relationships and multi-feature interactions, making it suitable for tasks where features have complex associations with MEDV. In the Boston dataset, the effect of DIS exhibits a pattern of diminishing returns: when the distance is less than 2, housing prices increase rapidly as distance decreases, but beyond 5, the effect weakens. A decision tree can accurately model this by setting thresholds. Similarly, the interaction between RAD (accessibility to highways) and TAX (property tax rate) shows that areas with good transportation but high taxes may have lower prices than areas with good transportation and low taxes; a decision tree can capture such complex relationships through hierarchical splits. The synergistic effect of CHAS (proximity to the Charles River) and RM, where riverfront houses with more rooms command significantly higher prices, can also be identified by decision trees, improving prediction accuracy.

Decision trees are insensitive to feature scales and do not require standardization, while effectively modeling nonlinear features such as NOX, AGE, and DIS. Experimental results show that they achieve the highest R^2, confirming their ability to capture complex patterns in the data. However, decision trees are prone to overfitting, particularly with limited samples, where they may fit noise (such as unusually high-priced houses in high-crime areas), reducing generalization ability. In practical applications, pruning or ensemble methods (such as random forests) are typically needed to optimize performance and this study does not explore such optimizations, which limits their applied potential.

3.2.3. Applicability of Bayesian regression models

Bayesian regression treats regression coefficients as random variables and updates posterior estimates using prior distributions. Its key advantage lies in providing information on prediction uncertainty, making it suitable for scenarios with small sample sizes, risk assessment needs, or probabilistic interpretation. If the Boston dataset has limited samples due to collection constraints, Bayesian regression can achieve more robust parameter estimates by specifying priors (such as assuming coefficients follow a normal distribution), helping to prevent overfitting. Its outputs include not only point predictions but also predictive distributions (such as mean and variance), allowing decision-makers to estimate, for example, that there is an 80% probability that MEDV falls between USD 20,000 and USD 25,000, providing probabilistic support for real estate pricing or mortgage risk assessment.

The model is relatively robust to minor noise, as prior distributions can constrain parameter ranges and reduce the influence of outliers. Experiments show that its MSE and MAE are close to those of linear regression, indicating good stability. However, it is still fundamentally based on a linear assumption and struggles to capture nonlinear patterns such as those in NOX and DIS, resulting in lower accuracy than decision trees. Moreover, the choice of prior distributions has a significant impact on performance: if set improperly (such as the non-normality of the true distribution), posterior estimates may be biased, reducing accuracy. This study does not optimize prior parameters, which limits the model's performance.

4. Discussion

4.1. Limitations

The Boston housing dataset is limited by the period in which the data were collected, and some features no longer accurately reflect the key factors influencing the current real estate market. For example, the dataset lacks modern housing characteristics such as smart home features or green building standards, which are important indicators for property valuation today, showing a clear temporal gap compared with more recent datasets like King County. In addition, the dataset does not include important macroeconomic variables that currently affect housing prices, such as interest rate fluctuations or mortgage policies, which may limit the models' ability to capture market dynamics.

Methodologically, this study does not implement a systematic hyperparameter optimization process. Most model parameters were set to their default values, without exploring the optimal parameter space using methods such as Grid Search. Research has shown that Grid Search, by exhaustively evaluating preset parameter combinations, can significantly improve model performance, though it is computationally expensive. Random Search, on the other hand, is more efficient in high-dimensional parameter spaces, and both approaches can achieve a 3-5% improvement in predictive accuracy for regression tasks. Additionally, there is room for improvement in feature engineering; advanced techniques such as LASSO regression for regularized feature selection or polynomial feature construction were not applied, potentially leaving redundant features that reduce model generalization. Similar studies have shown that removing irrelevant features via LASSO regression can reduce RMSE by 8-12% [8].

4.2. Future applications

In the field of real estate market prediction, a dual optimization framework of data updating and method upgrading can be established. The latest datasets from platforms like Kaggle (such as Bengaluru housing prices) should be used, as they include more granular location features and transaction information. In addition, unstructured data such as social media sentiment (such as school district reputation) and urban traffic flow should be integrated, with key influencing factors identified using methods like feature importance evaluation in random forests.

Model optimization can adopt a hybrid tuning strategy. For simpler models like linear and Bayesian regression, Grid Search can be used to optimize regularization parameters (such as L2 regularization coefficients), while for more complex models like decision trees, Random Search can improve tuning efficiency. Advanced approaches may incorporate ensemble algorithms like XGBoost, where parameter adjustment can achieve significant performance gains. Related studies show that a well-tuned XGBoost model can reach an R^2 of 0.83, significantly outperforming traditional regression methods.

The development of intelligent recommendation systems can focus on the integrated innovation of algorithms and user experience. On the algorithm side, personalized predictive models can be built based on user preference features (such as price sensitivity, preferred house layout), using techniques such as Grid Search-optimized LASSO regression for feature dimensionality reduction and weight adjustment. On the interaction experience side, VR-based house viewing data can be fed back into the model, and feature engineering can be used to create derived variables such as virtual viewing duration–purchase intent, improving the accuracy of recommendations.

In the field of environmental impact assessment, a multi-source data integration framework should be established. Environmental features such as air quality monitoring data (such as PM2.5 concentration) and noise pollution indices can be incorporated, with their association with housing prices quantified using correlation analysis methods similar to those in this study. Research shows that the correlation between green building features and housing prices can reach 0.62. The introduction of such features can enable models to provide a more comprehensive evaluation of livability.

In the field of financial product innovation, the risk assessment capabilities of models can be enhanced. Banks can leverage optimized regression models to develop submodules for predicting housing price volatility, using LASSO regression for feature selection to identify high-risk factors (such as high-crime areas) and provide data support for differentiated mortgage rate pricing. Additionally, the probabilistic outputs from Bayesian regression can be used to design hedging products against housing price declines, thereby improving the risk resilience of financial instruments.

5. Conclusion

Based on the Boston housing dataset, this study constructs three regression models, including linear regression, decision tree regression, and Bayesian regression, and compares them for predicting MEDV. Experimental results indicate that each model has its own strengths in terms of predictive accuracy, interpretability, and uncertainty modeling.

First, decision tree regression demonstrates the highest predictive accuracy, clearly outperforming the other two models. This indicates that the model can effectively capture nonlinear relationships and complex interactions between features and the target variable. However, decision trees also carry a high risk of overfitting, especially when proper pruning or parameter tuning is not applied. Therefore, in practical applications, it is recommended to combine cross-validation with ensemble methods (such as random forests or gradient boosting trees) to enhance the model's generalization ability.

Second, linear regression and Bayesian regression show similar overall performance, both demonstrating good stability and interpretability. Linear regression has a simple structure and fast training speed, making it suitable for initial modeling or as a baseline model. In contrast, Bayesian regression not only provides point estimates but also outputs predictive distributions, offering a unique advantage in scenarios requiring uncertainty assessment, such as risk evaluation or financial modeling. In addition, Bayesian regression is relatively robust to noisy data, making it suitable for situations with small sample sizes or unclear data distributions.

Notably, the Boston housing dataset used in this study is relatively old. Some features, including the absence of modern smart home and green building indicators, have limited relevance to the current real estate market. In addition, the models are not systematically hyperparameter-tuned, nor are advanced feature engineering techniques such as polynomial feature construction applied, which somewhat limits their generalization ability. Future studies can use more recent housing datasets that include macroeconomic variables, optimize model parameters via Grid Search, and apply LASSO regression to select key features, thereby further improving predictive performance and applicability in real-world scenarios.

References

[1] Harrison, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management, 5*(1), 81–102.

[2] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

[3] Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

[4] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. Springer.

[5] Harrison, D., & Rubinfeld, D. L. (1993). *Boston housing datase.UCI Machine Learning Repository*. https: //archive.ics.uci.edu/ml/datasets/Boston+Housing

[6] Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research, 13*, 281–305.

[7] Sablani, P. (2024). *King County house price prediction*. GitHub. https: //github.com/PrachiSablani/King-County-House-Price-Prediction

[8] Htsai, T. (n.d.). *House prices—Advanced regression techniques*. GitHub. https: //github.com/tifaniehtsai/Kaggle_Housing_Regression