Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 9
AIDML 2025

Comparing Linear Regression and Random Forest for Housing Price Prediction: Insights from the Boston Housing Dataset

Guoguo Chen *

Living Word Shanghai, Shanghai, China

* Corresponding Author Email: claraguoo@outlook.com

Abstract. Housing price prediction is a critical task in real estate and economic analysis, providing valuable insights for stakeholders such as homebuyers, sellers, and policymakers. This study focuses on the Boston Housing dataset, a benchmark with 505 samples and 14 features, to predict the median value of owner-occupied homes (MEDV) using Linear Regression and Random Forest Regression. Exploratory data analysis reveals non-linear patterns, such as the right-skewed distribution of MEDV (skewness = 1.11) and strong correlations with features like LSTAT (-0.74) and RM (0.70). The dataset was standardized and split into 80-20 training and testing sets for model evaluation. Results show that Random Forest outperforms Linear Regression, achieving an MSE of 7.58 and R^2 of 0.864 compared to 19.38 and 0.652, respectively. Feature importance analysis highlights LSTAT and RM as key predictors, emphasizing socio-economic and structural influences. While Random Forest excels in capturing non-linear relationships, Linear Regression offers interpretability for policy insights. However, the dataset's historical context and small size limit its applicability to modern markets, suggesting future research with larger, contemporary datasets and advanced models.

Keywords: Housing price prediction; Boston housing dataset; machine learning.

1. Introduction

Housing price prediction is the main character to analysis real estate and economic, and it also provides significant reference value for different people. Home buyer use it to judge whether he is affordable to buy the houses. Sellers use it to assess investment returns, and policymakers use it to plan cities and formulate economic policies. Understanding home rate changes, in city like Boston, seems especially complex and important due to its historical significance, socio-economic diversity, and industrial evolution. The Boston housing dataset, which is created by Harrison and Rubinfeld [1], is a classical case for regression studies. The dataset is collected in the 1970s, including 505 samples with 14 variables, such as per capita crime rate (CRIM), average rooms per dwelling (RM), proportion of residential land zoned for large lots (ZN), and percentage of lower-status population (LSTAT). The goal is to predicting the median value of owner-occupied homes (MEDV) in $1000s. Although the data is very old, its simple scale and abundant features make this data a common tool for predicting modeling. Since this data possesses both complexity and ease of operation, it is very suitable to explore regression technique.

The field of housing price prediction shows the multifaceted nature of research methods according to combining traditional methods of econometrics and innovative techniques of modern machine learning together. Harrison and Rubinfeld established the hedonic pricing model to separate house prices into contributions from observable attributes—structural (e.g., RM), environmental (e.g., nitric oxides concentration, NOX), and locational (e.g., distance to employment centers, DIS) [1]. Rosen claims that house price reflects buyers' preference for these characteristics and whether buyers are willing to pay [2]. This provides the theoretical basis for the Boston house price dataset. Later, Kain and Quigley incorporated social factors such as racial composition and low-income population into the model [3], which are directly related to variables such as LSTAT in the dataset. Traditionally, econometrics often uses linear regression to analyze housing prices. Mullainathan and Spiess contend that linear regression is suitable for causal inference because its coefficients are easy to interpret [4].

However, Fan and Lv points out that the assumption of linear regression is too simple to deal with some complex nonlinear relationships, such as the skewed distribution of the crime rate (CRIM) or the diminishing marginal effect of the proportion of low-income population (LSTAT) at high values [5]. Machine learning transforms the field of house price predictions. Breiman created random forests [6], which reduce errors and capture nonlinear relationships by integrate multiple decision trees. Liaw and Wiener verifies its validity [7]. However, Hastie et al. cautioned about the trade-off between interpretability and accuracy [8]. Bin and Kruse pointed out that complex models are prone to overfitting on small datasets (e.g., 505 samples) [9]. Other methods, such as support vector regression, geographically weighted regression, and spatial autocorrelation models, require larger or more detailed data. However, the dataset cannot provide.

This study compares linear regression and random forest regression. In addition, the study uses MSE and R^2 to evaluate performance and feature importance analysis to identify key predictors. The goal is to provide actionable insights for house price prediction. By contrast, the dataset's historical context and sample size is too small and lacks modern variables which may limit its applicability to contemporary markets.

2. Method

2.1. Dataset Preparation

This section summarizes the method of using Boston housing price dataset to predict housing price, including dataset preparation, model implementation, and evaluation metrics. This method combines traditional statistic technology with modern machine learning to compare their validity in regression tasks.

The study use the Boston housing price dataset from Kaggle which including 505 samples and 14 features with some categorical variables and a predictive goal. Table 1 provides detailed overview particularly.

Table 1. Descriptive Statistics of Key Features.

Feature	Mean	Std	Min	Max
CRIM	3.61	8.60	0.006	88.98
RM	6.28	0.70	3.56	8.78
LSTAT	12.56	7.14	1.73	37.97
NOX	0.55	0.12	0.38	0.87
PTRATIO	18.46	2.16	12.60	22.00
MEDV	22.59	9.20	5.00	50.00

The objective variable, MEDV, represents the median value of owner-occupied homes in $1000s as dependent variable of regression analysis. Exploratory data analysis is operated to understand the characteristic of data. The distribution of MEDV shown in Fig. 1 shows a right-skewed pattern (skewness = 1.11, kurtosis = 4.85) which manifest that the data is abnormal and contains latent value. This skewness indicates that there are fewer high-value houses which may challenge the assumption of a linear model of normality. Related analysis further reveals the significant relationship of features and objectives: LSTAT (percentage of lower-status population) shows a strong negative correlation with MEDV (-0.74), RM (average rooms per dwelling) has a positive correlation (0.70), while CRIM (per capita crime rate) displays extreme skewness (5.22) and heavy tails. These findings emphasis the non-linear relationship of data, the assumption of possibilities to challenge linear regression and necessitating a model capable of capturing these complexities.

Fig. 1 Housing price distribution (MEDV) (Picture credit: Original).

To prepare the data for modeling, this study applied some preprocessing procedures. First, the features were standardized to have a mean of 0 and a variance of 1, ensuring that variables with different scales (e.g., CRIM and RM) do not disproportionately influence the model. Second, the dataset was split into training and testing sets using an 80-20 ratio, resulting in 404 samples for training and 101 samples for testing. Setting random seeds as 42 to ensure repeatability of split. This dataset doesn't have the problem of missing values since they are pre-clean. However, the existence of the abnormal values distributing in MEDV are recorded because they may influence the property of model.

2.2. Machine Learning Models

This study employs two models implemented via scikit-learn to predict housing prices: Linear Regression and Random Forest Regression. These models were chosen to contrast a traditional statistical approach with a modern ensemble method, evaluating their ability to handle the dataset's non-linear characteristics.

Linear Regression. Linear Regression is a classical statistical method that assumes a linear relationship between the independent variables (features) and the dependent variable (MEDV) [10]. The model is defined as $\hat{y} = \beta 0 + \beta 1x1 + \beta 2x2 + \ldots + \beta nxn$, where $\hat{y}$ is the predicted MEDV, $\beta 0$ is the intercept, βi are the coefficients, and xi are the features. The model uses the ordinary least squares (OLS) method to train, which minimizes the sum of squared residuals between the predicted and actual values. No hyperparameters were using for Linear Regression, because it is a relatively simple model with no regularization in this implementation. The main advantage of Linear Regression is interpretability, as the coefficients βi directly shows the impact of each feature on MEDV. However, its assumption may limits its property because of the existing of non-linear pattern in data, such as the skewed distribution of CRIM and the complex relationship between LSTAT and MEDV.

Random Forest Regression. Random Forest Regression is an ensemble method which will solve the limitations of Linear Regression according to capture non-linear relationships and feature interactions. Random Forest constructs multiple decisional trees during training and outputs the average prediction of all the trees. It reduces variance and improving robustness compared to a single decision tree. The model is setting with 100 trees (n_estimators = 100) which is a common default that balances performance and calculating efficiency. Other hyperparameters, such as the maximum depth of trees and the minimum samples per split, keep the scikit-learn defaults to maintain the simplicity of preliminary of analysis. A key feature of Random Forest is its ability to provide feature importance scores according to the average reduction in variance across all trees when a feature is used for splitting. This is helpful for a deeper understanding of which variables, such as LSTAT or RM, most

strongly influence MEDV. Random Forest's flexibility makes it suited well to deal with non-linear effect and interaction in Boston housing price dataset. However, its complexities may increase the risk of overfitting on small data sets.

3. Results and Discussion

3.1. The Comparison of Model Performance

The performance of Linear Regression and Random Forest Regression was evaluated by using MSE and R^2 metrics, as shown in Table 2.

Table 2. Model Performance Metrics.

Model	MSE	R^2
Linear Regression	19.38	0.652
Random Forest	7.58	0.864

Linear Regression achieved an MSE of 19.38 and an R^2 of 0.652, indicating moderate predictive accuracy and explaining 65.2% of the variance in MEDV. In contrast, Random Forest significantly outperformed Linear Regression, with an MSE of 7.58 and an R^2 of 0.864. This represents a 61% reduction in MSE and a 32% improvement in R^2, demonstrating Random Forest's superior ability to capture the underlying patterns in the Boston Housing dataset. Fig. 2 illustrates the residuals versus predicted prices for both models.

Fig. 2 Residuals vs. predicted prices for Linear Regression and Random Forest (Picture credit: Original).

Linear Regression shows larger residuals, particularly for higher-priced homes, indicating their struggle with non-linear relationships. Random Forest, however, exhibits smaller and more evenly distributed residuals, confirming its robustness in handling complex patterns. Fig. 3 further supports this, plotting predicted versus actual prices.

Fig. 3 Predicted vs. actual prices for Linear Regression and Random Forest (Picture credit: Original).

Random Forest's predictions align more closely with the diagonal line (perfect prediction), while Linear Regression deviates more, especially for extreme values, reflecting its limitations with the dataset's non-linearities and outliers.

Table 3 provides the feature coefficients for Linear Regression, revealing the impact of each feature on MEDV.

Table 3. Linear Regression Coefficients.

Feature	Coefficient
CRIM	-0.11
RM	0.70
LSTAT	-0.74
NOX	-0.18
PTRATIO	-0.23

LSTAT has the strongest negative effect (-0.74), suggesting that a higher percentage of lower-status population significantly reduces housing prices. RM shows a positive coefficient (0.70), indicating that more rooms per dwelling increase the price. NOX (-0.18) and PTRATIO (-0.23) also have negative effects, reflecting the adverse impact of pollution and higher pupil-teacher ratios on housing values. CRIM's coefficient (-0.11) indicates a smaller but still negative influence of crime rate on prices. These coefficients align with economic intuition but highlight Linear Regression's limitation in capturing non-linear effects, as seen in the model's performance metrics.

3.2. Feature Importance Analysis

Random Forest's feature importance analysis, depicted in Fig. 4, provides deeper insights into the key drivers of housing prices.

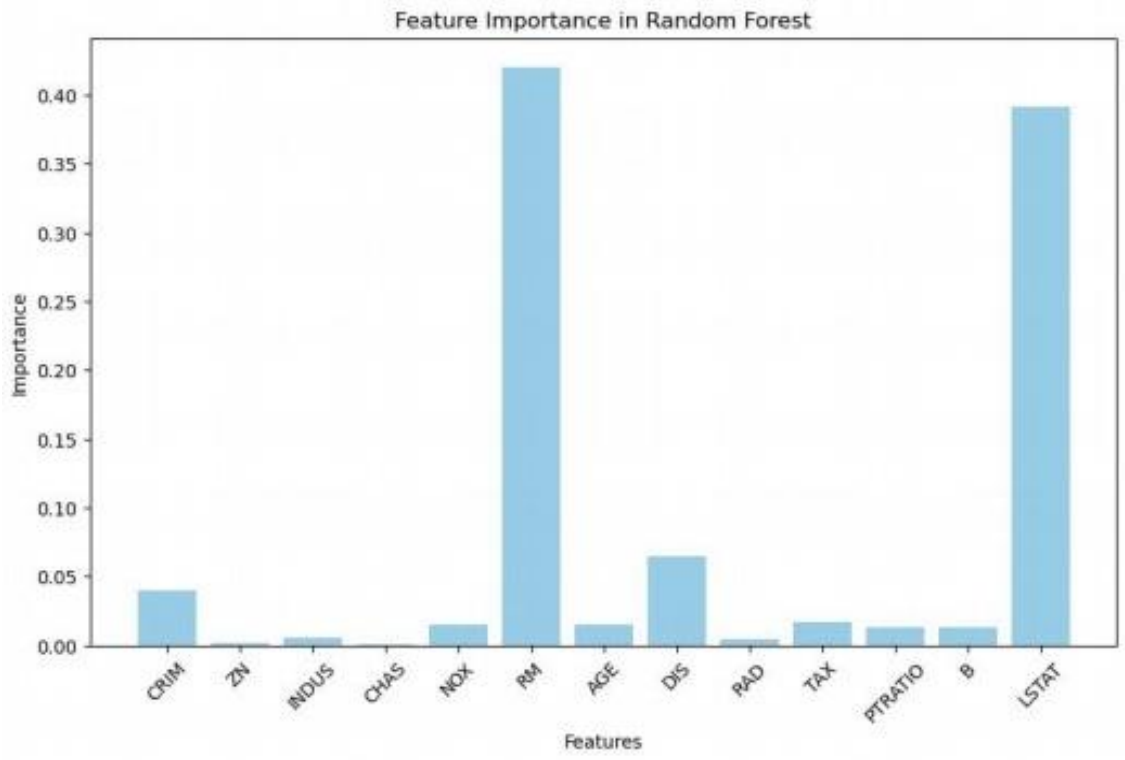

Fig. 4 Feature importance in Random Forest (Picture credit: Original).

LSTAT and RM emerge as the most influential features, with importance scores of approximately 0.40 and 0.35, respectively. This aligns with the correlation analysis in Section 2.1, where LSTAT and RM showed strong relationships with MEDV (-0.74 and 0.70). The high importance of LSTAT suggests that socio-economic factors play a critical role in determining housing prices in Boston, while RM underscores the value of structural attributes. Other features, such as NOX, DIS, and CRIM, have lower importance scores (around 0.05), indicating their relatively minor contributions. This analysis highlights Random Forest's ability to capture non-linear relationships and feature interactions, which Linear Regression struggles to model effectively. For instance, the interaction between LSTAT and RM—where the effect of room count may vary depending on the socio-economic context—is better captured by Random Forest, contributing to its superior performance.

The results demonstrate Random Forest's effectiveness in handling the Boston Housing dataset's complexities, making it a more suitable choice for real estate applications where predictive accuracy is crucial. However, Linear Regression's interpretability remains valuable for understanding direct feature impacts, which can inform policy decisions, such as addressing socio-economic disparities (LSTAT) or improving school quality (PTRATIO). The dataset's historical nature and small size limit the generalizability of these findings to modern contexts, necessitating future research with updated data.

3.3. Evaluation

Model performance uses two standard metrics for regression tasks to assess: Mean Squared Error (MSE) and R^2. MSE is defined as $MSE = (1/n) \Sigma (yi - \hat{y}i)^2$, where yi is the actual MEDV, $\hat{y}i$ is the predicted value, and the error is in \$1000s squared. A lower MSE indicates better predictive accuracy. R^2, the coefficient of determination, is calculated as $R^2 = 1 - (\Sigma (yi - \hat{y}i)^2 / \Sigma(yi - \bar{y})^2)$, representing the proportion of variance in MEDV explained by the model. R^2 ranges from 0 to 1. And a high value means a better fit. These metrics are calculated on the test set to evaluate the models' generalization performance. The function is to ensure whether it is fair to compare Linear Regression and Random Forest in capturing the underlying patterns of the Boston Housing dataset.

4. Conclusion

This study demonstrates that Random Forest outperforms Linear Regression in predicting Boston housing prices, achieving a significantly lower MSE (7.58 vs. 19.38) and a higher R^2 (0.864 vs. 0.652). Feature importance analysis reveals LSTAT and RM as the primary drivers, with scores of 0.40 and 0.35, respectively, underscoring the critical role of socio-economic and structural factors in determining housing values. Random Forest excels in capturing non-linear patterns and feature interactions, making it a more suitable choice for real estate applications where predictive accuracy is paramount. In contrast, Linear Regression provides valuable interpretability, offering insights into direct feature impacts that can inform policy decisions, such as addressing socio-economic disparities or improving educational resources. However, the Boston Housing dataset's historical context, small sample size (505 samples), and lack of modern variables—like technology-driven demand or gentrification effects—limit the generalizability of these findings to contemporary markets. Future research should focus on incorporating larger, more recent datasets to reflect current economic trends and explore advanced models like XGBoost or neural networks to further enhance predictive accuracy and applicability in real-world scenarios.

References

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

[2] Rosen S. Hedonic prices and implicit markets: Product differentiation in pure competition. Journal of Political Economy, 1974, 82(1): 34–55.

[3] Kain J F, Quigley J M. Housing markets and racial discrimination: A microeconomic analysis. Journal of Urban Economics, 1976, 3(3): 225–245.

[4] Mullainathan S, Spiess J. Machine learning: An applied econometric approach. Journal of Economic Perspectives, 2017, 31(2): 87–106.

[5] Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2014, 70(5): 849–911.

[6] Breiman L. Random forests. Machine Learning, 2001, 45(1): 5–32.

[7] Liaw A, Wiener M. Classification and regression by randomForest. R News, 2002, 2(3): 18–22.

[8] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer, 2009.

[9] Bin O, Kruse J B. Real estate market response to coastal flood hazards. Natural Hazards Review, 2019, 8(4): 121–132.

[10] Montgomery D C, Peck E A, Vining G G. Introduction to Linear Regression Analysis. John Wiley & Sons, 2021.