

Comparison and Analysis of Boston Housing Price Predictions Using Different Machine Learning Methods

Ruishan Liu *

Maple leaf international school, Xi'an, China

* Corresponding Author Email: 2196587957@qq.com

Abstract. In the contemporary real estate market, housing prices are fiercely competitive and influenced by various factors, such as the criminal activities existing in the community and the quality of the surrounding environment. However, inaccuracies in predicting house prices often lead to problems such as excessively high prices. The application of corresponding machine learning models plays a significant role in improving predictive accuracy in this domain. This study compares multiple models, including Ordinary Least Squares (OLS), Ridge regression, Lasso, Elastic Net, Random Forest, and Multilayer Perceptron (MLP), using a real-world housing price dataset. Model performance is evaluated through the accuracy, R^2 , RMSE, and MSE. The final results indicate that the Random Forest outperforms the other methods, achieving the highest predictive accuracy, while Ordinary Least Squares produces the lowest accuracy. Additionally, the study examines the applicability of these techniques across various scenarios and their potential to enhance the accuracy of housing price predictions.

Keywords: Linear Regression, Machine Learning, Boston House Prices Prediction.

1. Introduction

House price prediction has long been a central topic in economics, urban studies, and data science because of its profound impact on individuals, governments, and financial institutions. The housing market not only mirrors the economic vitality of a city but also influences social stability and long-term urban development. In this context, accurate housing price forecasts are crucial for developers, real estate investors, and households to make financial decisions. Estate enterprises aim to maximize profits from the sale of houses, while buyers purchase houses for personal residence or as investment assets. However, predicting house prices remains a challenging task because it is influenced by numerous factors, including the age and condition of the house, neighborhood crime rate, school quality, and others. Therefore, predicting house prices has become a crucial yet challenging issue in both academic and practical domains.

Among various datasets used for housing market research, the Boston Housing dataset has become one of the most widely studied benchmarks. In the 1970s, the housing prices in Boston were established, and various characteristics were incorporated into them. For instance, crime rates, employment centers, and so on. The dataset has been extensively applied to evaluate the effectiveness of different models and continues to serve as a fundamental reference in housing price prediction studies.

One of the earliest and most widely adopted methods for predicting Boston housing prices is linear regression. Its core principle is to discover the influence of different factors on the housing price prediction by minimizing the prediction error. However, the classical Ordinary Least Squares (OLS) method is prone to numerical instability and overfitting problems. To address these limitations, scholars have proposed two different improvement methods, Least Absolute Shrinkage and Selection Operator Regression (Lasso) and Tikhonov Regularization (Ridge Regression), which introduce L1 and L2 penalty terms [1]. Later, Elastic Net Regression combined the strengths of both methods, providing a more robust solution for complex datasets. Beyond linear regression, more advanced machine learning approaches have been introduced. Random Forest regression, as a nonlinear ensemble method, is capable of modeling complex feature interactions that linear models may

overlook. Similarly, neural networks provide a flexible framework that can approximate nonlinear relationships and adapt to intricate data structures.

This study compares six representative models based on the Boston housing price dataset, evaluates the predictive performance, analyzes the advantages and disadvantages of each model, and explores which method performs best under different application conditions, thereby providing a reference for actual housing price prediction.

2. Related work

Over the years, numerous studies have been conducted on predicting Boston housing prices, with various models proposed and tested. Different models demonstrated varying degrees of accuracy. At the beginning of the 20th century, economists Ragnar Frisch and Jan Tinbergen proposed the application of OLS for housing price prediction and other economic fields, including the analysis and modeling of predictions for stock price prediction and medical condition analysis. Subsequently, some researchers discovered the flaws of OLS and proposed relevant improvement methods. In 1970, the Ridge Regression model, improved by Hoerl and Kennard, was introduced [1]. By adding an L2 penalty term, the approach reduced some data errors but failed to screen variables. To address the shortcoming, Tibshirani proposed the Lasso model in 1996, which used an L1 as the penalty term to select and filter variables [2]. However, Lasso fails to effectively filter when variables are highly correlated. A further advancement is that Zou and Hastie improved OLS in 2005 and transformed it into the elastic net. The approach combined L1 and L2 to successfully overcome the shortcomings of the previous three models. It not only reduced errors but also enabled variable selection [3]. Subsequently, more scholars conducted related experiments, such as the work by Shariff and Duzan [4], which further demonstrated the advantages of Elastic Net.

Later, researchers began applying ensemble and nonlinear methods. Eze et al. [5, 6]. Incorporated the random forest model into the Boston housing price prediction. The final results demonstrated the advantages of this model, including data stability and low error, indicating its excellent predictive capabilities. This method prevented overfitting by using a random forest to handle noisy data and the bootstrap aggregation method. Meanwhile, some researchers pointed out that using MLP to predict house prices in Boston was feasible, as it could simulate the relationships between house prices and other influencing factors [7]. In these studies, precise data were obtained.

Despite these advancements, there had been no systematic comparison for the four models of elastic net, ridge regression, lasso regression, and ordinary least squares. Also, for the random forest and MLP included in the machine learning methods. This study aims to fill the gap by evaluating and comparing these six models using the Boston Housing Dataset to assess their predictive performance.

3. Methodology`

In this section, the operation methods of the six models are elaborated, as well as how different formulas are applied in relevant scenarios.

3.1. Ordinary Least Squares

The least squares method, initially proposed by Gauss, is mainly used in statistics and economics and serves as the most fundamental formula in linear regression [1]. Its operating principle is to estimate the regression coefficients by minimizing the sum of the squared errors between the predicted values and the actual values:

$$\hat{\beta}_{OLS} = \arg \min_{\beta} \sum_{i=1}^n (y_i - x_i^T \beta)^2 \quad (1)$$

In housing price prediction, if Y represents house prices, different X values correspond to different factors, such as the crime rate or average number of rooms [2]. The coefficients β indicate

the weight of different factors. The main purpose of OLS is to find a set of coefficients that minimizes the sum of these squared residuals.

However, OLS has several flaws and is unsuitable in complex scenarios. When multiple highly correlated independent variables are present, the OLS becomes unstable. Additionally, if the model is overly complex, the performance of OLS will significantly deteriorate, leading to overfitting. Therefore, OLS is only applicable to relatively simple models.

3.2. Ridge Regression

$$\hat{\beta}_{\text{Ridge}} = \arg \min_{\beta} \|Y - X\beta\|^2 + \lambda \|\beta\|_2^2 \quad (2)$$

Ridge regression was proposed by Hoerl and Kennard in 1970 [3]. It is used in various fields, such as macroeconomic modeling and stock market prediction. During the calculation, L2 regularization is added to the objective function, and L2 is derived through the basic OLS method.

Here, $\lambda \geq 0$ serves as a constraint, determining the intensity of regularization [3]. For instance, when predicting the housing prices in Boston, due to the high correlation between the ratio of students to teachers, the calculation result of OLS ranges from -5 to -50. However, Ridge Regression can narrow this range to -8 to -10, making the result more accurate and stable. It can prevent overfitting and control the complexity of the model and is very effective in dealing with the problem of multicollinearity among variables.

Nevertheless, the ridge regression model cannot compress the coefficients of weakly correlated variables to zero [6]. Therefore, it is impossible to automatically select the variables. Furthermore, selecting the optimal λ can be challenging and often requires cross-validation or other tuning methods.

3.3. Lasso Regression

The Lasso model was proposed by Tibshirani in 1996 [5]. and has been widely applied in finance, economics, and medical research, among other fields. It is also based on the OLS, with L1 serving as the regularization term and incorporating the absolute value penalty for the coefficients:

$$\hat{\beta}_{\text{Lasso}} = \arg \min_{\beta} \|Y - X\beta\|^2 + \lambda \|\beta\|_1 \quad (3)$$

Unlike ridge regression, which shrinks coefficients but cannot eliminate them, Lasso can directly reduce some coefficients to zero. For example, in predicting the housing prices of Boston, Lasso can identify the key factors, such as the crime rate or the student-teacher ratio, while shrinking less relevant factors to zero. This not only simplifies the model but also helps avoid overfitting. Thus, Lasso is suitable for datasets with numerous variables but limited relevant data.

However, when variables are highly correlated, the Lasso method is no longer applicable because it will only retain one and shrink the others to zero [5]. This selection causes uneven distribution, leading to unstable data. In such cases, Ridge is a better option.

3.4. Elastic Net

Elastic net was proposed by Zou and Hastie in 2005 [6,7]. Similar to ridge and lasso, it can be applied in industries and engineering, especially for dealing with multiple sensor correlations. The elastic network is not only based on OLS but also combines Ridge and Lasso, because it includes both L1 and L2:

$$\hat{\beta}_{\text{Lasso}} = \arg \min_{\beta} \|Y - X\beta\|^2 + \lambda \|\beta\|_1 \quad \hat{\beta}_{\text{EN}} = \arg \min_{\beta} \|Y - X\beta\|^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 \quad (4)$$

In the prediction of housing prices, if two factors are highly correlated but only weakly related to housing prices, the Elastic Net tends to simultaneously reduce the corresponding weights instead of randomly discarding one. This helps to increase the accuracy of housing price prediction, and it means

the Elastic Net can simultaneously filter variables and alleviate the problem of multicollinearity among variables.

Therefore, the method is applicable to scenarios where the variables are highly correlated. Because it can perform variable selection and ensure data stability at the same time, its comprehensive performance is stronger and more flexible, making it widely applied in various fields.

3.5. Random Forest

The Random Forest was officially proposed by Leo Breiman in 2001. It serves diverse purposes, including housing price prediction and applications in medical fields such as genetic data regression and classification. Random forests do not require a relationship (linear or otherwise) between the predicted target (dependent variable) and the input features (independent variables). This is the fundamental difference from the linear model. Here are the mathematical expressions:

$$\hat{y}_{RF}(x) = \frac{1}{B} \sum_{b=1}^B \hat{y}_{T_b}(x) \quad (5)$$

The prediction is obtained by aggregating the outputs of multiple decision trees. In the method, B refers to the specific number of trees used for training; the data comes from each tree of different bootstrap samples. The X represents the same input feature vector across all trees. Then, add up the predicted values of each tree, divide the total by the number of trees, and calculate the final average value, which is the final result.

This model has high robustness and generally provides more stable and accurate results compared with linear regression. However, training a large number of trees results in high memory consumption and long computation time. Moreover, the model's ensemble nature makes it difficult to determine which feature has a greater impact on the prediction [8].

3.6. Multilayer Perceptron (MLP)

The emergence of MLP has evolved through contributions from many researchers. Warren McCulloch and Walter Pitts proposed an early model in 1943 [9]. Subsequently, Frank Rosenblatt further developed the perceptron in 1958. In 1986, Rumelhart, Hinton, and Williams proposed the backpropagation algorithm, which enabled multi-layer networks to truly function [10].

The input layer, hidden layer, and output layer are these three components, constituting the classic MLP structure. The computation in a hidden layer is described by the following formula:

$$h_j = f\left(\sum_{i=1}^P w_{ij} x_i + b_j\right), j=1, \dots, H \quad (6)$$

Here, x_i represents the input factors, such as factors affecting housing prices like crime rate, student-teacher ratio, and so on, and w_{ij} denotes the weight connecting the input to the hidden neuron. b_j is the bias term, f is the activation function, and H is the number of hidden neurons [11]. The output layer computation is given by:

$$\hat{y} = v_1 h_1 + v_2 h_2 + \dots + v_H h_H + c \quad (7)$$

where v_H represents the corresponding weight connecting the hidden neuron to the output layer, and c is the output layer bias. $\hat{y}$ represents the final predicted value.

MLP offers many advantages, including the ability to capture complex relationship features, which is difficult to achieve with linear regression. It typically achieves superior performance on larger datasets compared to traditional datasets. However, MLP is not applicable in cases where the dataset is small. Additionally, the training process requires extensive hyperparameter tuning and architecture optimization, which can be computationally intensive and time-consuming [8].

4. Boston Dataset

This dataset was collected by Harrison, D., and Rubinfeld, D.L. in 1978. It encompasses various aspects of the real estate sectors in the Boston area of the United States from 1970 to 1979. The researchers aimed to use the data to analyze the relationships between various factors and house prices. Today, the Boston Housing Dataset remains a classic benchmark, whether in machine learning or linear regression.

The dataset contains 506 samples, each with 13 features and one target variable. The 13 characteristics, as shown in Table 1, include crime rate, land area exceeding 25,000 square feet, industrial and commercial land, riverfront location, nitrogen oxide concentration, average number of residential units, proportion of old houses, average distance to the center, and accessibility to the highway, as well as the ratio of teachers and the property tax rate for every 10,000 dollars [12]. The final variable is the target variable, namely the medv, which is predicted based on the aforementioned 13 features.

Table 1. Excerpt of Boston Housing Prices

crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	b	lstat	medv
0.09378	12.5	7.87	0	0.524	5.889	39.0	5.451	5	311	15.2	390.5	15.71	21.7
0.62976	0	8.14	0	0.538	5.949	61.8	4.708	4	307	21	396.9	8.26	20.4
0.63796	0	8.14	0	0.538	6.096	84.5	4.462	4	307	21	380.02	10.26	18.2
0.63739	0	8.14	0	0.538	5.834	45.8	4.499	4	307	21	395.62	8.47	19.9
1.05393	0	8.14	0	0.538	5.935	54.2	4.499	4	307	21	386.85	6.58	23.1
0.78420	0	8.14	0	0.538	5.990	58.7	4.258	4	307	21	386.75	14.67	17.5
0.80271	0	8.14	0	0.538	5.456	66.6	3.797	4	307	21	288.99	11.69	20.2
0.72580	0	8.14	0	0.538	5.727	96.1	3.797	4	307	21	390.95	11.28	18.2
1.25179	0	8.14	0	0.538	5.570	100	3.798	4	307	21	376.57	21.02	13.6
0.85204	0	8.14	0	0.538	5.965	85.9	4.012	4	307	21	392.53	13.83	19.6

5. Results

This part compares the performance of six models in predicting house prices in Boston using three evaluation metrics. They are R^2 , MSE, and RMSE. R^2 represents the explanatory power of the model, which is the coefficient of determination. The closer the value is to 1, the stronger the explanatory ability. MSE calculates the difference between the actual value and the predicted value. The smaller the difference between them, the higher the accuracy, that is, the mean square error. The final RMSE is the value obtained by taking the square root of MSE, that is, the root mean square error. The method for judging the accuracy is similar to MSE. The smaller the value, the more accurate the prediction. The experimental results are shown in Table 2.

Table 2. Performance Comparison of Models

Model	R^2	MSE	RMSE
Lasso ($\alpha=0.1$)	0.669	24.29	4.93
Ridge ($\alpha=1$)	0.668	24.31	4.94
OLS	0.65	25.66	5.07
Elastic Net ($\alpha=0.1$, $l1_ratio=0.5$)	0.656	25.2	5.02
Random Forest Regressor	0.892	7.9	2.81
MLP	0.895	14.33	3.79

Among linear models, the R^2 values of Lasso and Ridge are approximately equal to 0.67, and the error between the predicted values and the actual values is also relatively small. Elastic net shows slightly inferior results, which may be attributed to excessive regularization leading to loss of useful

information. OLS, lacking L1 and L2 regularization, shows larger errors as it is difficult to control irrelevant factors [1, 2].

Random Forest achieves the best overall performance, with an R^2 of 0.892 and the lowest MSE and RMSE values. The reason is that it has relatively high accuracy with low error rates and is less prone to overfitting and other issues [13]. This indicates that complex linear relationships between certain house prices and features can be effectively captured and resolved. The final results of MLP also demonstrate strong predictive ability compared to the linear model.

Overall, linear regression is suitable for simple model predictions but may not handle complex data well. Random forest shows the greatest advantage in housing price prediction, while the neural network, though slightly inferior to the random forest, still outperforms linear regression.

6. Discussion

Overall, the performance of these six models varies. The linear model includes OLS, elastic net, ridge regression, and lasso regression, but the final result is unsatisfactory overall. This is because the linear regression model is relatively simple and cannot detect or calculate more complex relationships in housing prices. MLP and Random Forests are capable of identifying such complex relationships. However, the most important thing is that MLP are more suitable for scenarios with large data volumes. Given that the current dataset is small, the performance of the MLP model was slightly inferior to that of the Random Forests [14].

This study has several limitations that should be acknowledged. The dataset employed is outdated and lacks the collection of more recent housing market information, which may affect the generalizability of the results. Furthermore, certain features were not adequately preprocessed. To address these issues and enhance future research, the following directions are recommended. First, more recent and diverse datasets should be collected, covering different regions beyond Boston, in order to draw more accurate conclusions. Additionally, a data cleaning protocol must be implemented, including systematic outlier detection and correction, to ensure data quality prior to modeling.

7. Conclusion

This study compares the OLS, Ridge, Lasso, Elastic Net, Random Forest, MLP, and random forest performance of these six models. On the same dataset for housing price prediction. The results show that regularized linear models such as Ridge and Lasso improve upon OLS in accuracy and stability, yet still underperform compared to random forests and MLP in capturing complex nonlinear information. Among linear models, Lasso facilitates variable selection but struggles with correlated features, while Ridge regression enhances stability but lacks feature selection capability. Based on these findings, this study provides a direction for housing price predictions, identifying Random Forest as the most suitable method based on our experiments. When large datasets are available, MLP can serve as a competitive alternative. In cases where these more complex models underperform, Lasso or Ridge regression offer viable fallback options due to their regularization properties.

For future work, integrating regularization techniques from linear models into ensemble and neural methods may help reduce overfitting and further improve predictive performance.

References

- [1] Arun K. Kuchibhotla, Lawrence D. Brown, and Andreas Buja. Model-free study of ordinary least squares linear regression. *Math S.T.*, 2018.
- [2] Maryam et al. Ordinary least squares (OLS) methods and issues: Theoretical foundations, challenges, and methodological remedies. *Pakistan Journal of Social Science Review*, 2025.
- [3] Hoerl, A. E., & Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 1970, 12(1): 55–67.

- [4] Roger W. Hoerl. Ridge regression: A historical context. *Technometrics*, 2020, 62(4):420-425.
- [5] Tibshirani, R. Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 2011, 73(3): 273–282.
- [6] Zou, H., Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 2005, 67(2): 301–320.
- [7] Hui Zou, Hao Helen Zhang. On the adaptive elastic-net with a diverging number of parameters. *Ann. Statist.*, 2009, 37(4): 1733-1751
- [8] Dr. Sweta R. Kumar et al. Predicting House Price with Deep learning: A comparative study of Machine Learning Models. *E-ISSN*, 2023, 2582-2160
- [9] McCulloch, W. S., Pitts, W. S. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 1943, 5(4): 115–133.
- [10] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. Learning representations by back-propagating errors. *Nature*, 1986, 323(6088): 533–536.
- [11] Chen, Z. Multi-level dense layer neural network model for housing price prediction. *arXiv preprint arXiv:2310.08133*, 2023.
- [12] Harrison, D., & Rubinfeld, D. L. Boston Housing dataset . U.S. Census Service, 1978.
- [13] Guoguo Chen. Comparing linear regression and random forest for housing price prediction: Insights from the Boston housing dataset. *Transactions on Computer Science and Intelligent Systems Research*, 2023.
- [14] Rumelhart, D. E., Hinton, G. E., Williams, R. J. Learning representations by back-propagating errors. *Nature*, 1986, 323(6088): 533–536.