# Research on housing prices prediction-take Boston as an example

#### **Boxuan Liu**

Dundee International Institute, Central South University, Changsha, 40086, China

7805220117@csu.edu.cn

Abstract. The problem of predicting Boston housing prices belongs to the field of artificial intelligence, specifically in the domain of regression problems, which is a crucial area of study in machine learning. This article aims to get an accurate prediction model of Boston housing prices by comparing four regression models based on the datasets derived from StabLib library. Multiple Linear Regression model, Random Forest Regression model, Extreme Gradient Boosting Regression model and Support Vector Machine Regression model are taken into consideration. Five evaluation index R-squared, adjusted R-squared, mean absolute error, mean squared error and root mean squared error are compared in terms of generalization ability of model. Eventually, Extreme Gradient Boosting Regression model was found to be the most effective model when predicting housing prices in Boston. The model has certain positive applications in real life, which can help government formulate real estate policies and people make wiser house purchasing strategies.

Keywords: Housing prices, regression model, prediction.

#### 1. Introduction

As an essential aspect of everyday life for the public, housing prices have consistently remained a matter of great importance. From 2004 onward, there has been a sustained nationwide escalation in housing prices in China, making it a central point of interest in the daily lives of individuals [1]. In China, mean house prices almost doubled from 2002 to 2010, indicating substantial fluctuations in house prices over the span of a few years [2]. The fluctuation in housing market prices has affected the daily consumption patterns of both urban and rural dwellers in China [3]. Also, housing price fluctuations significantly inhibit total factor productivity, which hindered economic development [4]. The similar thing happened in America. In the 80s and early 90s of the 20th century, housing prices in the United States maintained a steady upward trend, but Boston's housing prices fluctuated significantly, Boston's housing prices began to rise rapidly in 1983, and Boston's housing prices began to turn around in the third quarter of 1988, and it took four full years to get out of the shadow of the decline, which has had a certain impact on the economic development of Boston area [5]. Hence, housing prices fluctuations not only affect life of residents worldwide, but also hindered economic development in a certain degree. It shows that it is important to get an accurate housing prices prediction model to assist people in assessing the planned purchase of a house and government regulate housing prices for better economic development.

Many house prices prediction works have been carried out. Chen and Qing analyzed the influencing factors of housing prices in different communities in the Boston area in 1980, using quantile regression

<sup>@</sup> 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

as the basic method, and preliminarily explores the factors that affect the high or low prices of houses in different areas or communities in a region, besides the level of economic development. The use of quantile regression avoids the strict limitations of traditional ordinary least square (OLS) methods on data distribution characteristics and enables targeted research on different data at different quantiles [6]. Wang used the Boston housing price data, this study compares four methods: Levenberg-Marquardt, robust linear regression, Least Mean Squares, and tau. The focus is on analyzing the classic estimation methods and three robust estimation methods, studying the differences and advantages and disadvantages of the four methods [7]. But there exists a shortage of precision. Tian utilized a range of techniques to address the practical issue of predicting house prices in Boston, employing diverse methods for training and testing. It conducted a comprehensive comparison of diverse algorithm models, evaluating their performance from various perspectives and drawing conclusions about their effectiveness. It conducted a horizontal comparison of the advantages and weaknesses of various models and analyzed and summarized the differences in effectiveness [8]. Li and Guo considered the statistical diagnosis and local influence analysis of the single index mode based on data deletion and mode drift models. Simulation studies and results demonstrate the effectiveness and feasibility of the proposed method [9]. This article uses the typical and authoritative Boston housing price data as an example of the application of the semi-parametric varying coefficient quantile regression model and its two-stage estimation, demonstrating the model's good explanatory power in real economic issues [10].

In conclusion, this research compared different algorithmic models from multiple perspectives in order to provide an analysis and summary of their performance in predicting Boston house prices as a regression problem.

# 2. Methodology

## 2.1. Data source and description

The data utilized in this paper was sourced from the StatLib library, which is curated at Carnegie Mellon University. It contains 13 continuous attributes, 1 binary-valued attribute and 506 rows.

## 2.2. Variable selection

The Boston housing price data is relatively clean, so the preprocessing and feature engineering work will be relatively minimal. There are no missing values in the datasets. And the datasets don't contain attributes with low correlation, and there's no need to worry about multicollinearity. Eventually, all features are included. The data that the current paper analyzed contains 13 independent variables and one dependent variable (PRICE). The detailed information of this datasets is displayed in Table 1.

| Variable | Logogram        | Meaning  |
|----------|-----------------|--|
| CRIM     | x <sub>1</sub>  | Crime rate per capita in town  |
| ZN       | x <sub>2</sub>  | Percentage of residential land designated for lots exceeding 25000 sq.ft |
| INDUS    | x <sub>3</sub>  | Ratio of non-retail business acres per municipality                      |
| CHAS     | x <sub>4</sub>  | Charles River proximity (1 if tract borders river; 0 otherwise)          |
| NOX      | x <sub>5</sub>  | Concentration of nitric oxides (measured in parts per 10 million)        |
| RM       | x <sub>6</sub>  | Mean quantity of rooms per residence                                     |
| AGE      | x <sub>7</sub>  | Percentage of residences owned and constructed before 1940               |
| DIS      | x <sub>8</sub>  | Distances weighted by proximity to five employment hubs in Boston        |
| RAD      | x <sub>9</sub>  | Indicator for proximity to radial highways                               |
| TAX      | x <sub>10</sub> | Property tax rate per \$10,000 of assessed value                         |

 Table 1. Features interpretation.

Table 1. (continued).

| PTRATIO | x <sub>11</sub> | Ratio of students to teachers in each town                                      |
|---------|-----------------|---|
| В       | x <sub>12</sub> | 1000(Bk-0.63) <sup>2</sup> where Bk represents the percentage of blacks by town |
| LSTAT   | x <sub>13</sub> | The socioeconomic status of the residents                                       |
| PRICE   | Y               | Median value of homes owned by residents in \$1000s                             |

## 2.3. Method introduction

The paper compares four different methods: Multiple Linear Regression, Random Forest Regression, Extreme Gradient Boosting Regression, Support Vector Machine Regression. After analyzing the advantages and disadvantages of the four methods, the most suitable model was selected to construct a Boston housing prices prediction model based on the datasets.

In machine learning, Multiple Linear Regression is a statistical technique employed to forecast the linear correlation between several independent variables and a dependent variable. Within this method, the association is represented as a linear equation, in which the dependent variable is a weighted total of the independent variables, coupled with an intercept term. Random Forest Regression functions by creating numerous decision trees during the training process and produces the average prediction from each tree for regression assignments. Every tree is constructed using a random subset of the training data and a random subset of the features, aiding in diminishing overfitting and enhancing overall applicability. Extreme Gradient Boosting Regression operates by constructing a sequence of decision trees, with each subsequent tree rectifying the errors of its predecessor. A technique named gradient boosting is used to minimize the loss function by adding new models to the ensemble in a step-by-step fashion. Support Vector Machine regression aims to determine the optimal hyperplane that effectively segregates data points belonging to distinct classes within a high-dimensional space. It is applicable for both classification and regression assignments.

#### 3. Results and discussion

#### 3.1. Multiple linear regression

The mathematical formula for multiple linear regression is:

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{13} x_{13} + e$$
(1)

In the above formula:  $\beta_0$  is a constant term, and e is a residual term.

According to the heatmap, the independent variables do not exhibit highly correlated features among themselves and there are no features with relatively low correlation with the target variable (Figure 1). There is also no need to worry about multicollinearity.



Figure 1. Heatmap of correlation between features.

|--|

|          | В       | S.E.  | Т       | significance |  |
|----------|---------|-------|---------|--------------|--|
| Constant | 36.491  | 5.104 | 7.149   | 0.000        |  |
| X1       | -0.107  | 0.033 | -3.276  | 0.001        |  |
| X2       | 0.046   | 0.014 | 3.380   | 0.001        |  |
| X3       | 0.021   | 0.061 | 0.339   | 0.735        |  |
| X4       | 2.689   | 0.862 | 3.120   | 0.002        |  |
| X5       | -17.796 | 3.821 | -4.658  | 0.000        |  |
| X6       | 3.805   | 0.418 | 9.102   | 0.000        |  |
| X7       | 0.001   | 0.013 | 0.057   | 0.955        |  |
| X8       | -1.476  | 0.199 | -7.398  | 0.000        |  |
| X9       | 0.306   | 0.066 | 4.608   | 0.000        |  |
| X10      | -0.012  | 0.004 | -3.278  | 0.001        |  |
| X11      | -0.954  | 0.131 | -7.287  | 0.000        |  |
| X12      | 0.009   | 0.003 | 3.500   | 0.001        |  |
| X13      | -0.526  | 0.051 | -10.366 | 0.000        |  |

Regression outcomes are shown in the above (Table 2). The significance for the 11 independent variables did not exceed 0.002, which indicates that each of 11 independent variables has a notable influence on the dependent variable Y. To evaluate the model, visualization figure of actual prices compared to predicted prices is shown below (Figure 2).



Figure 2. Actual prices compared to predicted prices (multiple linear regression).

The visualized graph indicates that this is not a very good model in terms of intuitive effect. The deviation between predicted values and actual values is not slight enough to be ignored, which indicates that the model's fitting effect is not good (Table 3).

| Evolution          | Value  |
|--------------------|--------|
| Evaluation         | value  |
| R-squared          | 0.741  |
| Adjusted R-squared | 0.734  |
| MAE                | 3.273  |
| MSE                | 21.898 |
| RMSE               | 4.679  |

 Table 3. Model Evaluation (Multiple Linear Regression).

In order to evaluate the model and compare with other three models later more accurately, five data are shown in figure 3. R-squared score is 0.712. Adjusted R-squared score is 0.685. Mean absolute error is 3.867. Mean squared error is 30.068. Root mean squared error is 5.483.

#### 3.2. Random forest regression

Random Forest Regression is renowned for its robustness, flexibility, and capacity to manage highdimensional extensive datasets, making it a popular choice for predictive modeling in various domains.



Figure 3. Actual prices compared to predicted prices (random forest regression).

Figure 3 compares actual prices and predicted prices in the training set. If fitting all the points in the figure 3, the line is close to the line y=x. It turns out that the model has good generalization ability by visualization.

| Evaluation         | Value  |
|--------------------|--------|
| R-squared          | 0.823  |
| Adjusted R-squared | 0.805  |
| MAE                | 2.469  |
| MSE                | 18.619 |
| RMSE               | 4.315  |

Table 4. Model evaluation (random forest regression).

R-squared for Random Forest Regression is 0.822. Adjusted R-squared is 0.805. Mean absolute error is 2.469. Mean squared error is 18.619. Root mean squared error is 4.315 (Table 4).

## 3.3. Extreme gradient boosting regression

Extreme Gradient Boosting Regression is renowned for its high performance, flexibility, and excellent predictive capabilities, making it an essential tool for addressing regression problems.



Figure 4. Actual prices compared to predicted prices (XGBoost regression).

By visualization, it is evident that the Extreme Gradient Boosting model exhibits good fitting and generalization capabilities, notably outperforming the multiple linear regression model (Figure 4).

| Evaluation         | Value  |
|--------------------|--------|
| R-squared          | 0.849  |
| Adjusted R-squared | 0.835  |
| MAE                | 2.451  |
| MSE                | 15.716 |
| RMSE               | 3.964  |

Table 5. Model Evaluation (Extreme Gradient Boosting Regression).

For Extreme Gradient Boosting Regression model, R-squared is 0.849. Adjusted R-squared is 0.835. Mean absolute error is 2.451. Mean squared error is 15.716. Root mean squared error is 3.964 (Table 5).

## 3.4. Support vector machine

Benefits of Support Vector Machine Regression include efficacy with high-dimensional domains, robustness to overfitting, versatility in kernel selection, and suitability for non-linear data.



Figure 5. Actual prices compared to predicted prices (SVM egression).

Figure 5 shows the comparison of actual prices with prices predicted by Support Vector Machine Regression model. The scattered distribution of points in the graph indicates that the model has poor fitting and generalization capabilities.

| Evaluation         | Value  |
|--------------------|--------|
| R-squared          | 0.590  |
| Adjusted R-squared | 0.551  |
| MAE                | 3.756  |
| MSE                | 42.811 |
| RMSE               | 6.543  |

**Table 6.** Model Evaluation (Support Vector Machine Regression).

In fact, five evaluation index shown in table 6 reflects poor fitting and generalization capability of Support Vector Machine Regression model. R-squared is as low as 0.590. Adjusted R-squared is 0.551. Mean absolute error is 3.756. Mean squared error is as large as 42.811. Root mean squared error is 6.543.

## 3.5. Comparison of four models

In summary, table 7 gives a direct comparison of four regression models. Upon analyzing the performance of the four models, it is evident that Extreme Gradient Boosting Regression model demonstrates the highest R-squared and adjusted R-squared values, the lowest mean absolute error, mean squared error and root mean squared error. Therefore, Extreme Gradient Boosting Regression appears to be the most effective model among the four models.

|                            | $\mathbb{R}^2$ | Adjusted R <sup>2</sup> | MAE   | MSE    | RMSE  |
|----------------------------|----------------|-------------------------|-------|--------|-------|
| Multiple Linear Regression | 0.712          | 0.685                   | 3.867 | 30.068 | 5.483 |
| Random Forest Regression   | 0.822          | 0.805                   | 2.469 | 18.619 | 4.315 |
| Extreme Gradient Boosting  | 0.849          | 0.835                   | 2.451 | 15.716 | 3.964 |
| Support Vector Machine     | 0.590          | 0.551                   | 3.756 | 42.811 | 6.543 |

Table 7. Comparison Table for Four Models.

#### 4. Conclusion

The study applied four different regression methods to the datasets derived from the StatLib library. Upon analyzing R-squared value, adjusted R-squared value, mean absolute error, mean squared error

and root mean squared error, Extreme Gradient Boosting Regression was found to be the most effective model to predict Boston housing prices.

By the Boston housing prices prediction model using Extreme Gradient Boosting regression, the government can use the predictive results of the Boston housing price prediction model to formulate real estate policies such as rational planning of housing construction and controlling housing prices to promote sustainable urban development. For the general public and investors, the Boston housing price model can provide them with references when purchasing houses, helping them make wiser home purchasing and investment decisions. However, there are also some drawbacks. The model is based on historical data for prediction, and may not accurately predict the impact of future economic changes, policy adjustments, and other factors. And predictive results of the model are typically aimed at overall trends and may have errors in assessing the specific value of individual properties, unable to fully cover individual differences.

## References

- Wu Z K, Tang W G and Wu B 2007 Using the Priority Factor Method to Analyze the Impact of House Price Factors on Buyers' Orientation. Journal of Tianjin University of Commerce, 27(3).
- [2] Hu Q 2017 Analysis of housing price factors based on the SVAR model. Times Finance.
- [3] Yang D X and Zhang Z M 2013 An empirical study on incorporating housing price factors into China's CPI. Statistics&Information Forum, 28(3).
- [4] Li J, Chen Y and Guo W 2024 Will housing price fluctuations suppress total factor productivity growth? Empirical analysis based on 259 prefecture-level and above cities in China. Hainan Finance, 5, 3-21.
- [5] Yu X F 2002 Potential risks facing the current residential housing industry in Zhejiang from the perspective of international experience: Lessons from the substantial fluctuations in housing prices in Los Angeles and Boston from 1983 to 1993. Journal of Zhejiang University of Technology, 3.
- [6] Chen Z and Qing Q 2015 Factors affecting housing prices in different communities in Boston: Based on quantile regression analysis. Business, 30, 278-279.
- [7] Wang Y 2019 A review of research on Boston housing prices based on classical and robust methods. Market Weekly, 3, 40-43.
- [8] Tian R 2019 Boston housing price prediction based on multiple machine learning algorithms. China New Communication, 11, 228-230.
- [9] Li H, Zhu G and Guo Z 2017 Statistical diagnosis of single index mode and its application in the analysis of Boston housing prices. Statistical Research and Management, 6, 1091-1105.
- [10] Weng Y 2009 Semi-parametric varying coefficient quantile regression model and its two-stage estimation, Doctoral dissertation, Xiamen University.