Application of Machine Learning in Boston House Price Prediction

Yuanheng Zhang[1,a,*]

[1] *School of Microelectronics, Xi'an Jiaotong University, Xi'an 710049, China*
a. 196062145@mail.sit.edu.cn
** corresponding author*

Abstract: House prices are an important economic indicator for a country or region, rising house prices are often associated with economic growth and increased employment opportunities, while a decline in house prices may indicate an economic slowdown or other unfavorable factors. Governments and relevant agencies need to understand the dynamics of housing prices to formulate appropriate housing policies and plan urban development. As one of the major economic centers in the United States, the fluctuations in Boston's housing prices can reflect the local economic conditions and development trends. This paper selects data from Boston in 1970. Each record in this database describes a Boston suburb or town. The proposed method was evaluated using 5 metrics. By comparison, the paper ends up with the result that XGBoost works best out of the four regression models.

Keywords: machine learning, Boston house price, XGBoost

1. Introduction

1.1. Background Information and Restrictions

Like the less transparent industry in our ecosystem, housing prices change day by day and are sometimes exasperated rather than based on evaluation. Forecasting shelter prices based on real elements is the primary component of this research. This document aims to do our estimations on account of each basic metrics which is taken into consideration while setting the price.

Real estate holds significant importance as an investment field, wherein precise house price prediction facilitates informed investment decisions for real estate developers, investors, and homebuyers [1].

Predicting house prices has become a burning subject of research recently. Due to the sharp increase in housing demand, it is vital to develop end-user expectations for receiving comprehensive home information. However, due to price and demand instability, it is increasingly challenging for users to make up their mind. By incorporating a predictive analytics system into house pricing, multiple stakeholders can be helped in effectively planning and making decisions. Actually, the size of the housing market has increased significantly, and in order to reduce the demand needs, it is essential to introduce the new approach, which may help solve the issues [2].

House prices serve as crucial indicators of the economy and financial markets. Studying the housing price problem is conducive to improving the governance ability of the government. In the past research, the research on house prices has always been limited to urban areas, but there is a lack of research on house prices in different regions. However, due to its unique attributes, house prices are closely related to different regions of the city. Therefore, it is very necessary to study the problem of house prices in the regional stage [3]. This enables them to formulate relevant policies, implement regulatory measures, and exercise macro-control to ensure market stability and sustainable development.

Numerous factors affect the price of homes, including geographic location, housing characteristics, macroeconomic indicators, demographics, and other variables. The interplay between these factors is intricate and non-linear, making accurate housing price prediction a complex task. It necessitates the comprehensive analysis of extensive data and their interrelationships.

There are also some limitations in traditional house price predicting methods. The linear models commonly employed in traditional methods assume a linear relationship between housing prices and their characteristics. However, housing prices are influenced by a multitude of complex factors, rendering this simplistic linear assumption insufficient for capturing the intricate nonlinear relationships accurately.

By analyzing and interpreting conclusions from data patterns with the aid of computers and mathematical frameworks, machine learning is the study of how computer systems are used and developed to learn and adapt without specific guidance [4]. Furthermore, these metrics are highly valuable in the modern era as they have the capability to capture intricate nonlinear connections between various factors that impact housing prices. Their significance extends to various domains, ranging from everyday tools such as search engines and email filters to more intricate applications like predicting customer behavior or, in the case at hand, forecasting housing prices.

1.2. Target and Problem

Conventional methods of predicting house prices rely on comparing the cost of a house with its actual sale price. However, forecasting becomes challenging due to the wide array of factors that impact the housing market. A shortage of information that could increase the effectiveness of the real estate market can be filled by the creation and accessibility of numerous house price prediction models [5].

Machine Learning can be effectively used for house price prediction, which has b the potential to be a powerful tool to improve the accuracy and effectiveness of house price forecasts.

So, this paper takes Boston as a typical example. Boston is one of the oldest and most significant historical and cultural cities in the country. It is also the city with the highest level of education in the country and the hub of higher education, healthcare, and investment money. In bustling urban business districts, investors are always eager to pursue real estate, and changes in housing prices and rents depend on many factors. If a ML model can correctly forecast the price of a Boston home given a set of characteristic data, that is the issue that this project will try to solve.

2. Several Factors

2.1. Data sources

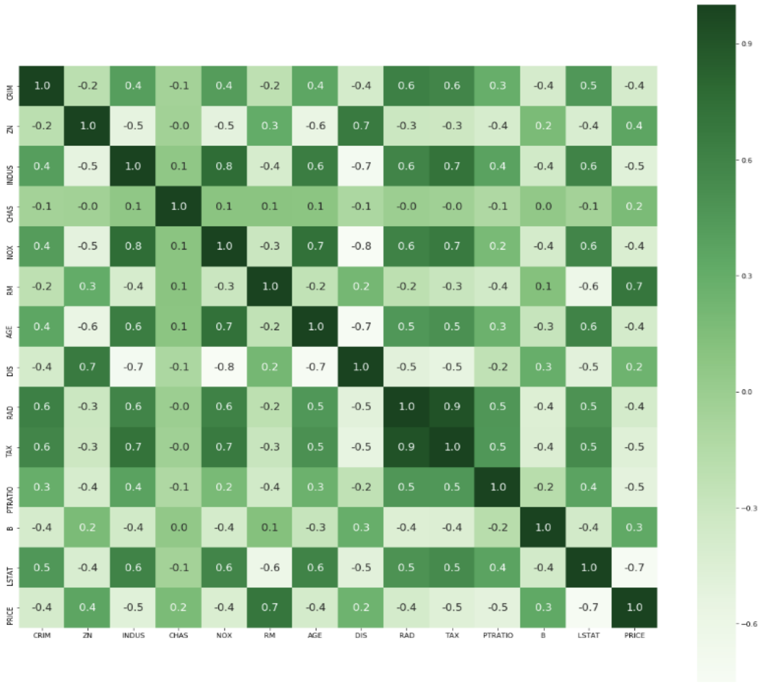
To form the machine learning template with Boston housing data, this paper will use the data selected from Boston in 1970. Every record in this data set corresponds to a Boston city or suburb. 506 rows, 13 attributes (characteristics), and a target column (price) are provided.

2.2. Features

This paper first checked the dataset for errors, missing values, outliers or duplicate records and clean it up.

After that, the correlation between the features need to be found out by using heatmap. Heatmaps really provide an intuitive way to visualize patterns and relationships in data. By using color gradients, they represent different values or levels of a variable, allowing patterns and trends to be easily identified. Heatmaps are particularly useful for identifying clusters, hotspots, or areas of high and low values within a dataset.

Heatmaps also allow for quick comparisons between different variables or categories. By placing multiple heatmaps side by side or using color scales, it becomes easier to identify similarities, differences, or correlations between different sets of data. This comparative analysis can reveal insights that may not be immediately apparent from looking at the raw data (See Fig. 1).



	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	PRICE
CRIM	1.0	-0.2	0.4	-0.1	0.4	-0.2	0.4	-0.4	0.6	0.6	0.3	-0.4	0.5	-0.4
ZN	-0.2	1.0	-0.5	-0.0	-0.5	0.3	-0.6	0.7	-0.3	-0.3	-0.4	0.2	-0.4	0.4
INDUS	0.4	-0.5	1.0	0.1	0.8	-0.4	0.6	-0.7	0.6	0.7	0.4	-0.4	0.6	-0.5
CHAS	-0.1	-0.0	0.1	1.0	0.1	0.1	0.1	-0.1	-0.0	-0.0	-0.1	0.0	-0.1	0.2
NOX	0.4	-0.5	0.8	0.1	1.0	-0.3	0.7	-0.8	0.6	0.7	0.2	-0.4	0.6	-0.4
RM	-0.2	0.3	-0.4	0.1	-0.3	1.0	-0.2	0.2	-0.2	-0.3	-0.4	0.1	-0.6	0.7
AGE	0.4	-0.6	0.6	0.1	0.7	-0.2	1.0	-0.7	0.5	0.5	0.3	-0.3	0.6	-0.4
DIS	-0.4	0.7	-0.7	-0.1	-0.8	0.2	-0.7	1.0	-0.5	-0.5	-0.2	0.3	-0.5	0.2
RAD	0.6	-0.3	0.6	-0.0	0.6	-0.2	0.5	-0.5	1.0	0.9	0.5	-0.4	0.5	-0.4
TAX	0.6	-0.3	0.7	-0.0	0.7	-0.3	0.5	-0.5	0.9	1.0	0.5	-0.4	0.5	-0.5
PTRATIO	0.3	-0.4	0.4	-0.1	0.2	-0.4	0.3	-0.2	0.5	0.5	1.0	-0.2	0.4	-0.5
B	-0.4	0.2	-0.4	0.0	-0.4	0.1	-0.3	0.3	-0.4	-0.4	-0.2	1.0	-0.4	0.3
LSTAT	0.5	-0.4	0.6	-0.1	0.6	-0.6	0.6	-0.5	0.5	0.5	0.4	-0.4	1.0	-0.7
PRICE	-0.4	0.4	-0.5	0.2	-0.4	0.7	-0.4	0.2	-0.4	-0.5	-0.5	0.3	-0.7	1.0

Figure 1: Heatmap.

3. Method and Result

3.1. Important Parameters

When evaluating regression models, it is important to consider these metrics together. A high R-squared and adjusted R-squared, along with low MAE, MSE, and RMSE, cite a model that, with the fewest possible errors in forecasting, illustrates most of the variance in the dependent variable. However, when analyzing these indicators, it is also crucial to also take into account the specific context and needs of the analysis.

3.2. Linear Regression

LR is a ML method that belongs to supervised learning. It derives its name from the concept of a linear relationship between variables. Given that there is a linear relationship between them, it demonstrates how the dependent variable is dependent on the independent variables. On a graph, the connection between them can be shown as a single line [6].

Linear regression is a statistical method employed to look into the connection between a response variable and one or multiple predictor variables. It operates on the assumption of a linear connection between these variables, implying that the dependent variable can be expressed as a linear combination of the independent variables.

This paper trained a linear regression model using the training dataset and the created a data frame, which shows the coefficients of the linear regression model on each attribute of the training dataset. Intercept and coefficient are also extracted in order to make it more convenient to give further analysis and explanation.

This paper calculated the metrics for model evaluation on the prediction results of the training dataset. For training data and test data, the results are shown in Table 1 and Table 2, respectively:

Table 1 : Results of linear regression for training data.

	Value
R2	0.746
Adjust R2	0.736
MAE	3.089
MSE	19.073
RMSE	4.367

A dispersion diagram was utilized to demonstrate the disparities between the actual prices and the expected values. The x-axis represents actual prices, and the y-axis represents expected prices. This dispersion diagram visually represents the relationship between real and forecast prices. The dots along the diagonal line indicate a close correspondence between the actual and expected values, while the deviations from the line indicate differences between them (See Fig. 2).

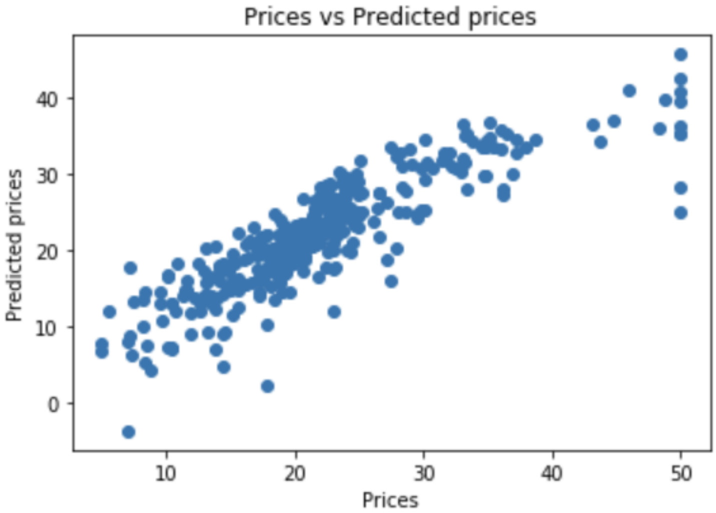


Figure 2 : Predicted prices and actual prices under linear regression.

Table 2:Results of linear regression for test data.

	Value
R2	0.712
Adjust R2	0.685
MAE	3.859

Table 2: (continued).

MSE	30.054
RMSE	5.482

3.3. Random Forest Regressor

Many different machine learning techniques are currently used to successfully tackle a wide range of data categorization problems. To ensure the accuracy of the classification decisions, they must "fine-tuning" the values of their parameter values [7].

The Random Forest algorithm, introduced by Breiman, enhances the predictive accuracy of a model by consolidating numerous regression trees. By leveraging this approach, the algorithm can effectively summarize and combine the outputs of these trees, resulting in improved prediction performance. The Random Forest technique can reflect the interaction between factors and makes it simple to calculate the nonlinear effects of variables. The nonlinear relationships between the dependent and independent variables can therefore be expressed [8].

This subject train the model using the training dataset by calling the method with the training features and the corresponding target variable. For training data and test data, the results are shown in Table 3 and Table 4, respectively:

Table 3:Results of random forest for training data.

	Value
R2	0.966
Adjust R2	0.965
MAE	0.943
MSE	2.552
RMSE	1.598

These metrics help assess how well the random forest regressor model fits the training data and how accurately it predicts the target variable. A scatter plot is created to visualize the relationship between the actual prices and the predicted prices (See Fig. 3).

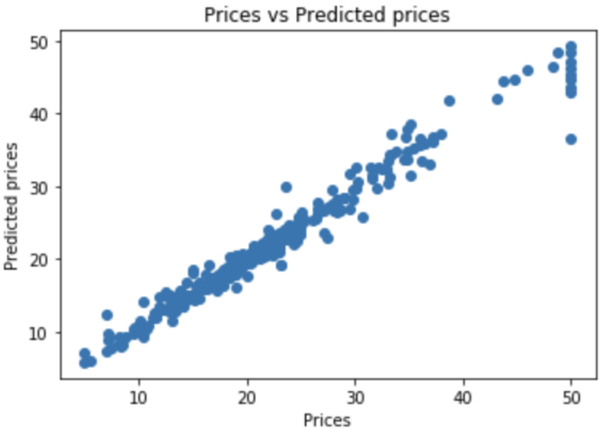


Figure 3: Predicted prices and actual prices under random forest.

Table 4：Results of random forest for test data.

	Value
R2	0.864
Adjust R2	0.851
MAE	2.568
MSE	14.195
RMSE	3.768

3.4. XGBoost Regressor

When predicting property prices, the XGBoost algorithm can adequately capture the nonlinear relationship [9]. Nevertheless, the parameters of XGBoost must be chosen carefully because they affect the algorithm's capacity to learn and generalize. After training the model by using the training dataset, the XGBoost Regressor is able to make predictions on fresh, unforeseen data. For training data and test data, the results are shown in Table 5 and Table 6, respectively:

Table 5:Results of XGBoost for training data.

	Value
R2	0.970
Adjust R2	0.969
MAE	1.137
MSE	2.231
RMSE	1.494

A scatter plot is created to visualize the relationship between the actual prices and the predicted prices under XGBoost regressor (See Fig. 4).

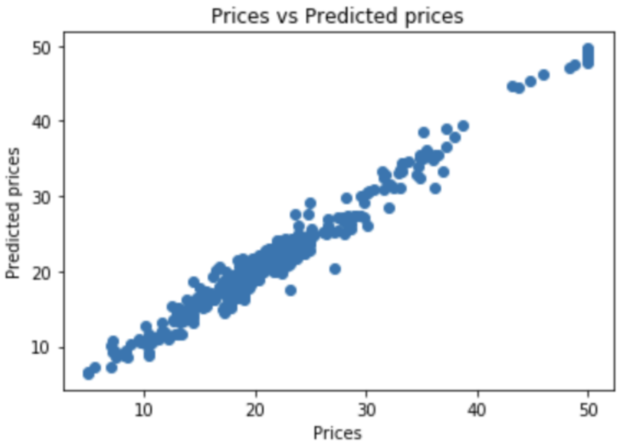


Figure 4：Predicted prices and actual prices under XGBoost.

Table 6: Results of XGBoost for test data.

	Value
R2	0.849
Adjust R2	0.835
MAE	2.451

Table 6:(continued).

MSE	15.716
RMSE	3.964

3.5. SVM Regressor

SVM-r is a method used to connect input data to a single output. Its purpose is to minimize the estimation error by reducing the discrepancy between the predicted and actual fault distances. As a regression tool, the SVM has a clear mathematical definition [10]. For training data and test data, the results are shown in Table 7, Table 8 and Fig. 5, respectively:

Table 7: Results of XGBoost for training data.

	Value
R2	0.642
Adjust R2	0.628
MAE	2.937
MSE	26.954
RMSE	5.192

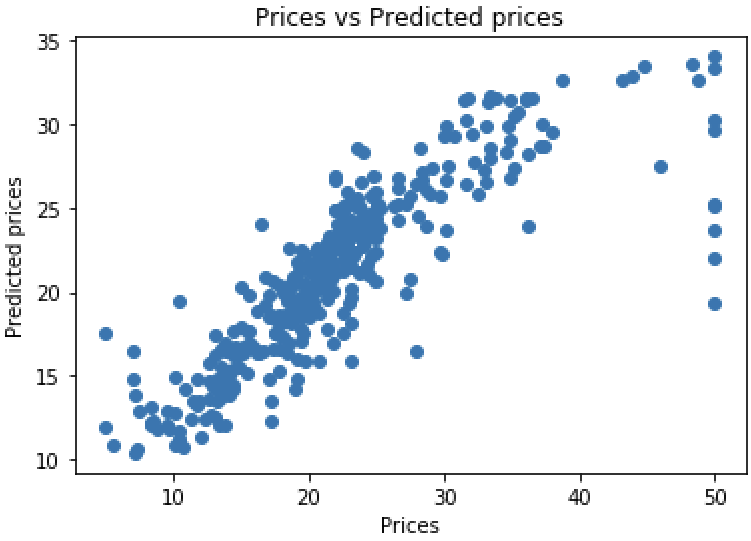


Figure 5: Predicted prices and actual prices under SVM.

Table 8: Results of XGBoost for test data.

	Value
R2	0.590
Adjust R2	0.551
MAE	3.756
MSE	42.811
RMSE	6.543

4. Evaluation and Comparison

Finally, a dataframe is created which contains the names of different models and their corresponding R-squared scores. The R-squared scores are multiplied by 100 to represent them as percentages (See Table 9).

Table 9: R-squared Score of the four models.

	Model	R-squared Score
1	Random Forest	86.406
2	XGBoost	84.948
0	Linear Regression	71.218
3	Support Vector Machines	59.001

In this analysis, the Boston dataset utilized consists of a relatively small number of instances, specifically 506. Despite its size, significant insights can still be derived from this dataset. To assess these models' performance, the data were divided into a training package comprising 70% of the instances and a test package comprising the remaining 30%. Various metrics, including R-squared, RMSE, MAE, and MSE, were employed to assess the model's accuracy and predictive capability. Hence Random Forest works the best for this dataset with a R-squared Score of 86.406.

5. Conclusion

This study trained and tested four different machine learning models using a dataset of 506 observations. In order to assess how well these models predicted the future, five separate measures were implemented. In the end, XGBoost proved to be the best of the four models. The discipline of Boston house price forecasting has benefited greatly from this study's numerous significant contributions. It highlights the importance of utilizing machine learning techniques for accurate and reliable predictions. The study demonstrates that these models outperform traditional linear regression models by capturing non-linear patterns and incorporating a wide range of features.

However, it is essential to acknowledge the limitations of this study. The findings are dependent on the quality and representativeness of the dataset used. Furthermore, the selected machine learning models may have certain limitations in terms of interpretability and transparency. Future research could focus on incorporating additional datasets, exploring other advanced machine learning algorithms, and developing hybrid models that combine the strengths of different approaches.

In summary, this study illustrates the effectiveness of ML techniques for predicting housing prices in Boston. By continuously improving and refining these models, we can improve our understanding of housing market dynamics and develop more accurate and reliable forecasting models.

References

[1] *Varma, A., Sarma, A., Doshi, S., Nair, R.: House price prediction using machine learning and neural networks. In 2018 second international conference on inventive communication and computational technologies, 1936-1939 (2018).*

[2] *Cekic, M., Korkmaz, K. N., Müküs, H., Hameed, A. A., Jamil, A., Soleimani, F.: Artificial intelligence approach for modeling house price prediction. In 2022 2nd International Conference on Computing and Machine Intelligence, 1-5(2022).*

[3] *Li, Y., Zhang, R., Wang, J., Shi, J.: Channel exchange network model for house price prediction. International Conference on Machine Learning and Intelligent Systems Engineering, 96-99 (2022).*

[4] *Bai, S.: Boston house price prediction: machine learning. International Conference on Intelligent Computing and Signal Processing, 1678-1684 (2022)*

[5] *Muralidharan, S., Phiri, K., Sinha, S. K., Kim, B.: Analysis and prediction of real estate prices: a case of the Boston housing market. Issues in Information Systems 19(2), 109-118 (2018).*

[6] *Chandu, P., Devi, N. B.: Improved Prediction Accuracy of House Price Using Decision Tree Algorithm over Linear Regression Algorithm. International Conference on Science Technology Engineering and Mathematics, 1-6 (2023).*

[7] *Demidova, L., Ivkina, M.: Approach to determining the boundaries of the search range for the number of trees in the random forest algorithm. In 2020 9th Mediterranean Conference on Embedded Computing, 1-4 (2020).*

[8] *Dong, P., Peng, H., Cheng, X., Xing, Y., Zhou, X., Huang, D.: A random forest regression model for predicting residual stresses and cutting forces introduced by turning in718 alloy. International Conference on Computation, Communication and Engineering, 5-8 (2019).*

[9] *Sheng, C., Yu, H.: An optimized prediction algorithm based on XGBoost. In 2022 International Conference on Networking and Network Applications, 1-6 (2022).*

[10] *Correa-Tapasco, E., Perez-Londoño, S., Mora-Florez, J.: Setting strategy of a SVM regressor for locating single phase faults in power distribution systems. In 2010 IEEE/PES Transmission and Distribution Conference and Exposition: Latin America, 798-802 (2010).*