Predicting Boston Housing Price Using Machine Learning Models

Hanqiu Ding

School of Arts and Sciences, Brandeis University, 415 South St, Waltham, MA 02453, The United States
hanqiuding@brandeis.edu

Abstract. In response to the ongoing rise in living expenses, especially the rise of rental prices, in Boston, people are increasingly inclined to explore alternative options for long-term financial savings. Compared to having an expense with no ownership, purchasing real estate is one of the top optimal options as it serves as a long-term investment. Given this big environment, utilizing predictive machine learning models (MLM) is a way for people to figure out the factors that influence Boston housing prices. Although most of the existing research utilized advanced MLM techniques or a combination of several regression models to increase the accuracy of the prediction, a few studies focus on basic MLM. In this paper, three traditional models are utilized to predict the features that affect Boston housing prices: the Multiple Regression Model (MLR), Random Forest (RF), and eXtreme Gradient Boosting (XGBoost). The paper uses the Interquartile Range (IQR) method to remove the outliers in the dataset and handles the missing values by checking the amount and dropping them. Last but not least, the research uses R-squared, adjusted R-squared, cross-validated R-squared, and Root Mean Square Error (RMSE) as performance indicators of those models. The result shows that XGBoost has the best performance among the others.

Keywords: Multiple Regression Model, Random Forest, eXtreme Gradient Boosting.

1 Introduction

As a big city with renowned universities and firms, Boston attracts a large flux of people every year. Those individuals, such as professors, students, and technicians, need to consider the living cost of Boston when deciding to relocate to Boston. As the living costs, especially the rental prices, in Boston keep rising every year, people tend to seek alternative ways to save money in the long run. For financial consideration, owning a real estate is an optimal choice to save money. Rather than giving a certain amount of money to others consistently, paying a fixed monthly mortgage to own a property serves as a suitable option for long-term investment. However, buying houses is a complicated process as there are various factors for people to be taken into consideration. Hence, an in-depth analysis of the features of the house and a reliable

B. Siuta-Tokarska et al. (eds.), *Proceedings of the 2024 2nd International Conference on Management Innovation and Economy Development (MIED 2024)*, Advances in Economics, Business and Management Research 300,
https://doi.org/10.2991/978-94-6463-542-3_51

prediction of the price of the house provide insights for people who aim to buy houses in Boston.

There are a portion of researchers who use different models to make predictions. Ye employs the Ridge Model (RM) and the RF Model improved by the genetic algorithm to predict Boston housing prices and provides a comparative analysis. The study found that the RF Model improved by the genetic algorithm has a better performance of the prediction compared to the prediction of the RM [1]. Sanyal employs the Boston home dataset Using Ridge, Lasso, Polynomial, and Simple Linear regression (SLR), to develop an advanced automated MLM. The experiment shows that Lasso has a better performance than the others [2]. Zhao establishes a Linear Regression (LR) after the Box-Cox transformation and the Lasso to predict Boston housing prices through the R language. The outcome suggests that non-linear regression models should be investigated further as LR may not be the optimal choice [3]. Based on the dataset of house properties and crime data, Muralidharan makes predictions about the appraised values of residential properties through the use of MLM-like artificial neural networks and decision trees. The study found that neural networks have better performance [4].

While most of the studies focus on SLR, only a few of them focus on MLR. This paper uses MLR and compares it with other models. In this paper, 13 independent variables that affect Boston housing prices are taken. By comparing MLR, RF, and XGBoost, this paper seeks to explore the model that makes the best prediction.

2 Data Description

The Boston housing dataset originates from the University of California, Irvine Machine Learning Repository and is recollected by Jangir [5]. There are 14 features in the dataset, containing 13 independent variables and 1 dependent variable. Among all independent variables, there are two categorical variables; the remaining features are continuous. There are 506 entries in the dataset. However, one of the variables has five missing values. Hence, further work requires handling the missing data and identifying outliers to ensure the accuracy and applicability of the model. The coding part of this paper refers to the code from Rutecki [6]. This paper further adds the method for calculating the test error of the training data and provides a horizontal comparison of the test error of different regression models for both testing and training data.

3 Methodology

3.1 Outliers Removal

This paper calculates the skewness of all independent variables, and the skewness of them is either positive or negative. This means that the distributions of those variables are not symmetric, which may not be in a normal distribution. Since most of the variables are not in a normal distribution, the data preprocess includes detecting and removing outliers in the dataset. The IQR Method is introduced to find the lower and

upper boundaries of the dataset. Any value exceeding the range is considered an outlier.

Upper Boundary: $Q3 + 1.5$ (IQR)

Lower Boundary: $Q1 - 1.5$ (IQR)

where Q1 and Q3 correspond to the first and third quartiles. IQR equals to Q3 minus Q1.

Since some of the regression models, such as LR, are sensitive to outliers, removing outliers from the dataset ensures a better performance of the prediction of the models.

3.2 Handling Missing Value

There are five missing values in the column 'rm'. Since the missing value accounts for a small proportion of the data size, dropping the missing values in corresponding rows has little effect on the dataset and the prediction.

3.3 Measurements of the Test Error

R-squared.

The equation of the R-squared is as follows:

$$R^2 = \frac{ESS}{TSS} \tag{1}$$

ESS stands for Explained Sum of Squares, the square of the total amount that differs between the mean and the estimated value, $\sum_{i=1}^{n}(\tilde{Y}i - \bar{Y})^2$.

TSS stands for Total Sum of Squares, the square of all the variations between the mean and the actual value, $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$.

R-squared, also known as the coefficient of determination, is a statistical metric that quantifies the goodness of fit of a regression model to the data. It is always in the range of 0 to 1. It shows that a specific proportion of the variability of y close to its mean may be explained by the model. The greater the percentage of Y's fluctuation that the model can account for, the closer it gets to 1.

Adjusted R-squared.

The equation is the following:

$$R^2 = 1 - \frac{n-1}{n-k-1}\,(1 - R^2) \tag{2}$$

The number k in the model denotes the count of regressors, whereas n is the number of rows in the dataset.

R-squared rises in multiple regression when an additional regressor or independent variable is included in the model. The mechanical rise in R-squared resulting from the addition of regressors is adjusted using adjusted R-squared.

Cross-validated R-squared.

Since R-squared is generated based on the entire dataset, it may overestimate the performance of the model. Cross-validated R-squared deals with the overestimation by taking the average of R-squared for each round of cross-validation.

RMSE.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y})^2}{n}} \tag{3}$$

RMSE measures the mean discrepancy between the observed and anticipated values. The model fit and target value prediction are both improved by a reduced RMSE value.

3.4 MLM

MLR.

The equation for the MLR is as follows:

$$Yi=\beta 0 + \beta 1X1 + \beta 2X2+\ldots+\beta kXk + Ui \tag{4}$$

Yi is the dependent variable, and X1, .., and Xk are k independent variables, which are also called regressors. β0 is the intercept of the model, β1, ..., βk are the coefficients of the regressors. Ui is the error term of the model.

The MLR Model is a measurement exploring the relationship between a response variable and several explanatory variables.

RF.

RF is a method for combining various tree predictors so that each tree in the forest has the same distribution and is reliant on the values of a randomly selected vector [7]. Random Forest increases tree variation and prevents tree correlation by allowing trees to develop from several training data subsets generated through a process called bagging. As a result, some data may be utilized repeatedly during the training, whereas some data may not be utilized at all. This increases the stability and improves the prediction accuracy [8].

XGBoost.

A machine-learning system for tree boosting known as XGBoost offers flexibility in handling large datasets. The majority of tree learning algorithms currently in use are either limited to processing dense data exclusively or call for certain procedures to handle certain circumstances, such as categorical encoding. Unlike those tree learning algorithms, XGBoost handles all sparsity patterns uniformly [9]. The advantages of XGBoost are its speed and model functionality. It applies to cache optimization, out-of-core computing, distributed computing, and parallelization [10].

4 RESULTS

Table 1. Test error for testing and training data

	R-squared		Adjusted R-squared		Cross-validated R-squared		RMSE	
MLM	Training	Testing	Training	Testing	Training	Testing	Training	Testing
XGBoost	1	0.9043	1	0.8870	0.8143	0.8143	0.0006	2.5847
RF	0.9485	0.8683	0.9465	0.8445	0.7653	0.7653	1.6610	3.0318
MLR	0.7251	0.7674	0.7142	0.7254	0.6876	0.6876	3.8389	4.0288

As shown in Table 1, compared to the other two models, XGBoost has a better performance in predicting the dependent variable. MLR model has the lowest value for three types of R-squared, which means that there might be a non-linear relationship between independent variables and the dependent variable. It also shows that the linear model does not fit with the dataset even if more regressors are involved. Even though test errors for Random Forest are smaller than those for MLR, XGBoost performs better than Random Forest.

5 Conclusion

This paper uses MLR, RF, and XGBoost to predict the housing price in Boston. The results of this investigation show that XGBoost works the best among the other two models as its three types of R-squared are all close to 1 and it has the smallest RMSE. This result demonstrates that the variation in housing prices can be explained the most by XGBoost. On the other hand, the MLR model has the worst performance as it has the lowest R-squared and highest RMSE. The MLR's performance indicates that the dataset is not a suitable fit for the MLR. Also, the relationship between explanatory variables and the response variable might be non-linear. Compared with the results of all models, such as Lasso and Ridge, in Sanyal's paper, the R-squared of XGBoost in this paper has the highest value, which means this model explains a higher percentage of the variation in Boston housing price compared to all others. However, compared to the R-squared of RF improved by genetic algorithms in Ye's paper, XGBoost has a lower R-squared, which means it still needs to improve. This difference suggests that more advanced techniques such as algorithms or a combination of regression models should be introduced to the prediction procedure to ensure higher accuracy of the model.

References

1. Ye, L.: Comparison of ridge regression and GA-RF models for Boston house price prediction. International Journal of Mathematics and Systems Science, 6(4), 377-382 (2023).
2. Sanyal, S., Shrestha, S., Satpathy, S.: Boston house price prediction using regression models. In 2022 2nd International Conference on Intelligent Technologies (CONIT), pp. 1-6. IEEE (2022).
3. Zhao, R.: Analysis of the correlation between housing price data in Boston based on the regression method (2020).
4. Muralidharan, S., Shamsuddin, A.: Analysis and prediction of real estate prices: A case of the Boston housing market. Issues in Information Systems, 19(2), 109-118 (2018).
5. Jangir, A.: Boston Housing Dataset. Kaggle. https: //www. kaggle. com/ datasets/ arunjangir245/boston-housing-dataset (2022).
6. Rutecki, M.: Regression models evaluation metrics. Kaggle. https: //www. kaggle. com/ code/marcinrutecki/regression-models-evaluation-metrics/notebook#5.-Data-pre-processing (2022).
7. Breiman, L. Random forests. Machine learning, 45(1), 5-32 (2001).
8. Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., Chica-Rivas, M.: Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees, and support vector machines. Ore Geology Reviews, 71, 804-818 (2015).
9. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. arXiv preprint arXiv: 1603.02754 (2016).
10. Zaki, J., Hanif, M., Syed, I. A.: House price prediction using hedonic pricing model and machine learning techniques. Concurrency and Computation: Practice and Experience, 34(27), e7342 (2022).