Machine Learning based Comparative Study on House Price Prediction Task

Ruodong Hu[1, *]

[*] Corresponding author: dc02755@um.edu.mo

[1]Faculty of Science and Technology, University of Macau, Macau, 999078, China

Abstract: Because it enables buyers and sellers to make informed decisions about purchasing and selling real estate, house price prediction is important. Additionally, it aids banks and other financial organizations in determining the worth of a property when taking into account a mortgage or loan application. It also aids real estate agents and brokers in pricing homes appropriately. The use of machine learning and neural networks to anticipate home prices has been extensively studied. In one publication, an approach is put out for predicting housing prices with real inputs utilizing a variety of regression techniques. In order to examine the precision of the four models—linear regression, random forest regressor, XGBoost regressor, and support vector machine regressor—in predicting Boston home prices, no study has been done. The importance of this study is to evaluate how well these four models forecast Boston home values.

Keywords: house price, prediction, model, regression

1. INTRODUCTION

1.1 Background

The aim of house price prediction is to determine how much a property will be worth in the future based on various aspects, such as location, size, condition, features, market trends, and economic conditions. House price prediction is beneficial for both buyers and sellers, as it can help them make wise choices and bargain better. House price prediction is also significant for investors, lenders, insurers, and policymakers, as it can impact the profitability, risk, and stability of the housing market [1,2].

House price prediction is a difficult task, as it involves working with complex and dynamic data that may be incomplete, noisy, or inconsistent. Furthermore, house price prediction is affected by many aspects that are hard to measure or quantify, such as consumer preferences, expectations, emotions, and behavior. Therefore, house price prediction needs advanced methods and techniques that can capture the nonlinear and heterogeneous relationships between the variables that influence house prices.

Machine learning is a field of artificial intelligence that allows computers to learn from data and make predictions without explicit programming. Machine learning has been extensively used for house price prediction in recent years, as it can deal with large and diverse data sets and uncover hidden patterns and features that may not be evident to human experts. Machine

MSEA 2023, May 26-28, Nanjing, People's Republic of China

DOI 10.4108/eai.26-5-2023.2334435

learning can also adjust to changing market conditions and provide precise and timely forecasts. [3]

1.2 Purpose

House price prediction is the process of using data analysis methods to estimate the sale price of a house based on its features and market condition. The goal of house price prediction is to help buyers and sellers make reasonable trading strategies, avoid blindly following the trend or missing opportunities. The necessity of house price prediction is because the real estate market has characteristics such as complexity, uncertainty and non-linearity, which make house prices affected by various factors and difficult to grasp accurately. Therefore, using scientific methods and techniques to predict house prices can improve decision efficiency and accuracy. The main contributions of this work is presented as follows:

1. Use linear regression, random forest regressor, XGBoost regressor, and support vector machine regressor predict and analyze the vacation in Boston.

2. Check the residuals corresponding to linear regression, random forest regressor, XGBoost regressor, and support vector machine regressor respectively.

3. Analyze the accuracy of linear regression, random forest regressor, XGBoost regressor, and support vector machine regressor and compare them.

2. METHOD

2.1 Dataset description

It is a CSV file which is related to house price in Boston. Each entry in the database provides information on a town or suburb of Boston. The Boston Standard Metropolitan Statistical Area (SMSA) in 1970 served as the source of the statistics. The definitions of the properties are extracted from the UCI Machine Learning Repository which can be seen in table 1.

Table 1. Annotation of each term

Abbreviation	The meaning of abbreviation
CRIM	Rate of crime per person in a given town
ZN	what proportion of homes are on lots that are more than 25,000 square feet
INDUS	percentage of land each municipality uses for non-retail companies
CHAS	auxiliary variable for the Charles River (= 1 if the tract borders the river, 0 otherwise)
NOX	The level of nitric oxides in the air
RM	The typical number of rooms per residence
AGE	The percentage of residences built for their owners before 1940
DIS	Weighted distances to five Boston-area job locations
RAD	Accessibility score for radials on the highway
TAX	$10,000 is a rate of total-assessment property tax
PTRATIO	school-to-student ration by town
B	1000(Bk−0.63)^2,where Bk is the number of black people living in each

	municipality
LSTAT	Lower population status as a percentage
MEDV	Median owner-occupied house value in the $1,000 range

2.2 Algorithm

2.2.1 Linear Regression

Definition: Statisticians utilize a method known as linear regression to describe the connection between a dependent variable and one or more independent variables. The traits or situations that influence the outcome are known as independent variables, while the outcome or response that we are attempting to predict or explain is known as the dependent variable. An equation of the form $y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \varepsilon$, where y is the dependent variable, may be used to express the linear or straight-line connection between the variables in linear regression [4]. $X_1, \ldots, X_n$ are the independent variables, $\beta_0, \ldots, \beta_n$ are the coefficients that measure how much each variable affects y, and ε is the error term that captures the variability not explained by the model [4].

$$\hat{y}(\omega, x) = \omega_0 + \omega_1 x_1 + \ldots + \omega_p x_p \tag{1}$$

Set the vector $\omega = (\omega_1, \ldots, \omega_p)$ as coefficient and ω_0 as intercept.

2.2.2 Random Forest Regressor

Definition: The Random Forest regressor is a machine learning technique that performs regression problems by using numerous decision trees. It is a particular kind of ensemble learning technique that integrates the forecasts of various models to increase accuracy and decrease overfitting [5]. A subset of the data and features, selected at random with replacement, is used to train each decision tree in a random forest. As a result, there will be more variation and less association between the trees. The average of all the trees' predictions is used to get the final prediction of a random forest regressor. Random forest regressor can handle nonlinear relationships, missing values, outliers, and high-dimensional data [5].

2.2.3 XGBoost Regressor

Definition: XGBoost regressor is a machine learning algorithm that uses extreme gradient boosting to perform regression tasks. Gradient boosting is a method for sequentially constructing an ensemble of decision trees, with each tree attempting to fix the flaws of the one before it. XGBoost improves upon gradient boosting by adding regularization terms to prevent overfitting, parallelizing the tree construction for faster computation, and handling missing values and sparse data. XGBoost regressor can handle nonlinear relationships, high-dimensional data, and complex interactions among features [6].

$$Obj^{(t)} \approx \sum_{i=1}^{n} [g_i f_i(x_i) + \frac{1}{2} h_i f_i^2(x_i)] + \sum_{i=1}^{t} \Omega(f_i) \tag{2}$$

2.2.4 SVM Regressor

The SVM regressor is a method for machine learning that performs the support vector machine (SVM) to do regression tasks. SVM is a technique that locates a hyperplane that divides the data points into various classes. SVM regressor adapts this idea to fit a function that minimizes the error between the predicted and observed values within a certain margin, called epsilon-insensitive loss. SVM regressor can handle both linear and nonlinear relationships by using different kernel functions, such as polynomial, radial basis function (RBF), or sigmoid. SVM regressor can also handle high-dimensional data and outliers [7].

$$\begin{aligned} &\min_{a,b} = \frac{1}{2} \| a \|^2 \\ &s.t. y_i (a^T x_i + b) \geq 1, i = 1, 2, ..., m \end{aligned} \tag{3}$$

3. RESULT

3.1 Correlation Analysis

The correlation between the features and plotting the heatmap of correlation between features of house in Boston is shown in figure 1.

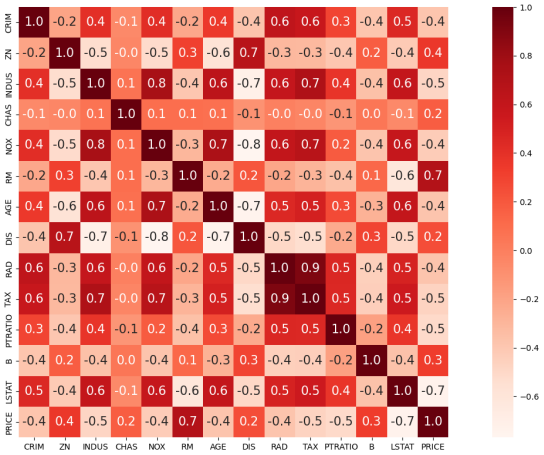
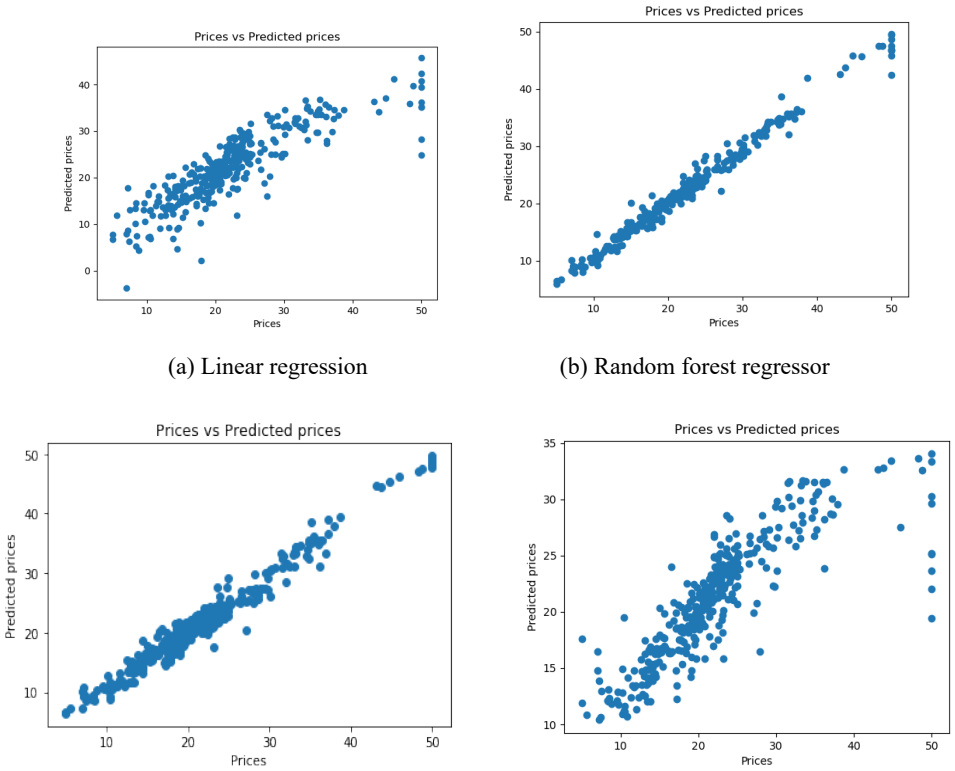


Figure 1. Heatmap of correlation between features of house

3.2 Visualizing the differences between actual prices and predicted values

Four different models are applied to predict the price of the houses in Boston, which are linear regression, random forest regressor, XGBoost regressor and support vector machine regressor. The comparison results of actual prices and predicted values are shown below.



(a) Linear regression (b) Random forest regressor

(c) XGBoost regressor (d) SVM regressor

Figure 2. Visualizing the differences between actual prices and predicted values

If the predicted value is closer to the actual value, the points on the graph are more concentrated near the graph of the proportional function. The assumption that XGBoost regressor is the most accurate model can be established from figure 2.

3.3 Checking residuals

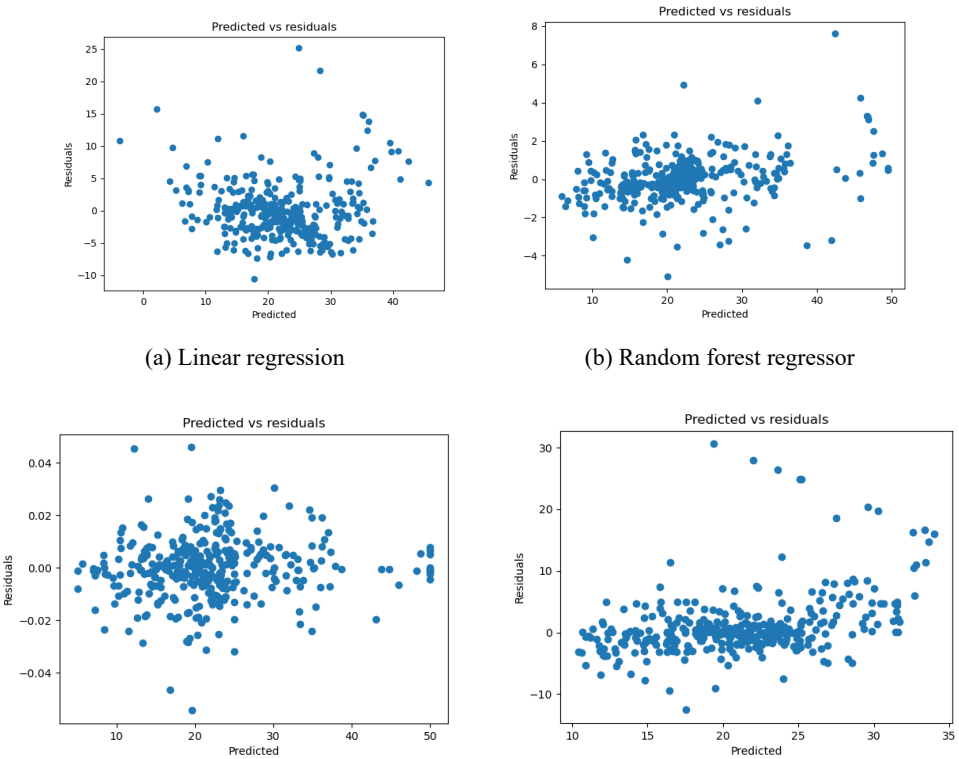
The discrepancies between the values that were seen and those that were expected in a regression analysis are known as residuals. They gauge how well the data points match a regression line. Residuals are computed as follows: Residual = Observed value - Predicted value [8].

Depending on whether the observed value is above or below the regression line, residuals can either be positive or negative. If the observed value is greater than the anticipated value, the residual is said to be positive; if it is lower, the residual is said to be negative. If the residual is 0, the observed value and the anticipated value are exactly same.

The closer a residual is to zero, the better the fit of the regression line to that data point. The sum of all residuals in a regression analysis is always zero. This is because the regression line

minimizes the total squared residuals, which is why it is also called the least-squares regression line.

Residuals are useful for checking the assumptions of a regression model, such as linearity, homoscedasticity, independence and normality. They can also be used to detect outliers, influential points and non-linear patterns in the data [8].



(a) Linear regression (b) Random forest regressor

(c) XGBoost regressor (d) SVM regressor

Figure 3. Checking residuals

The closer the fluctuation of residuals is to 0, the more accurate the model is. The residuals of XGBoost regressor, random forest regressor, linear regression and support vector machine regressor fluctuate within a range of (-0.04,0.04), (-3,3), (-7,7) and (-10,10) respectively generally. According to figure 3, it can be convinced that XGBoost regressor is the most accurate.

3.4 Evaluation and comparison of all models

Table 2. Evaluation results

Model	R-squared Score
XGBoost	85.7995
Random Forest	82.8694

Linear Regression	71.2182
Support Vector Machines	59.0016

The goal of this analysis was to compare four different models for predicting house prices in Boston using Python. The models were XGBoost, random forest, linear regression and support vector machines (SVM). The R-squared score, which gauges how well a model matches observable data, was utilized as the performance parameter to assess the models. A better fit and a smaller prediction error are both indicated by a higher R-squared value.

The results given in table 2 showed that XGBoost had the highest R-squared score of 86, followed by random forest with 83, linear regression with 71 and SVM with 59. This means that XGBoost was the best model for predicting house prices in Boston among the four models tested. It was able to capture most of the variation in the data and produce accurate predictions. Random forest was also a good model, but slightly less effective than XGBoost. Linear regression was a moderate model, but had a higher prediction error than XGBoost and random forest. SVM was the worst model, as it had a low R-squared score and a high prediction error.

There are several possible reasons for the differences in performance among the models. One reason is that XGBoost and random forest are ensemble methods that combine several weak learners to create a powerful learner. They can handle complex nonlinear relationships and interactions among the features, as well as reduce overfitting and variance [9,10]. Linear regression and SVM are single learners that assume a linear or nonlinear relationship between the features and the target variable. They can be affected by outliers, multicollinearity, heteroscedasticity and high dimensionality [4,7].

Another reason is that XGBoost and random forest have hyperparameters that can be tuned to optimize their performance [5,9]. For example, XGBoost has parameters such as learning rate, max depth, subsample and colsample by tree that control the learning process and the complexity of the trees. Random forest has some parameters split that control the number and size of the trees and the randomness of the features [10]. Linear regression and SVM have fewer or no hyperparameters to tune, which limits their flexibility and adaptability [7,11].

A third reason is that XGBoost and random forest are robust to missing values and outliers in the data. They can handle missing values by splitting on them or imputing them with mean or median values [6,9]. They can also handle outliers by using robust splitting criteria such as entropy or gini index. Linear regression and SVM are sensitive to missing values and outliers in the data. They require complete data or imputation methods such as mean or median imputation or k-nearest neighbors imputation. They can also be affected by outliers by producing large residuals or errors [11,12].

This analysis demonstrated that XGBoost was the best model for predicting house prices in Boston among the four models tested. It had the highest R-squared score of 86 and the lowest prediction error. Random forest was also a good model, but slightly less effective than XGBoost. Linear regression was a moderate model, but had a higher prediction error than XGBoost and random forest. SVM was the worst model, as it had a low R-squared score and a high prediction error. The differences in performance among the models were due to their

ability to handle complex nonlinear relationships, interactions, missing values, outliers and hyperparameters in the data.

4. CONCLUSION

House price prediction is the process of analyzing and forecasting the future trend and change of house prices using methods such as statistics, machine learning, artificial intelligence, etc., based on historical data and influencing factors. House price prediction has important necessity and prospect, mainly reflected in the following aspects: House price prediction can help the government formulate reasonable housing policies, balance supply and demand, and prevent bubbles or crashes in the real estate market. House price prediction can help developers, investors, banks and other market participants grasp market dynamics, optimize resource allocation, improve profitability and risk management. House price prediction can help buyers, sellers, renters and other consumers make rational decisions, save costs and improve their quality of life.

With the development of technology and the increase of data, the methods of house price prediction are also constantly improving and innovating. Currently, some common methods of house price prediction include: Linear regression model: using multiple linear regression to analyze various factors affecting house prices (such as area, location, facilities etc.), and establishing mathematical equations to estimate future house prices. Spatial model: taking into account the spatial correlation and heterogeneity between different regions, using spatial weight matrix or spatial lag model to capture spatial effects and improve prediction accuracy. Spatiotemporal model: combining time series analysis and spatial analysis using spatiotemporal autoregressive models or spatiotemporal Kalman filters to describe the patterns of house prices changing over time and space. Machine learning model: learn complicated data features by adaptive optimization and generalization utilizing nonlinear algorithms including artificial neural networks, support vector machines, decision trees, random forests, and gradient boosting trees.

Future integrated house price forecast will be more accurate and intelligent in real-time as additional dimensions and types of data (such as social media satellite images sensor data, etc.) are obtained and put to use, including big data cloud computing Internet of things. At the same time some challenges risks such as data quality security interpretability and so forth need to be paid attention to. In summary house price prediction is a field with broad significance huge potential there are still many spaces worth exploring innovating in the future.

REFERENCES

[1] Ahtesham, Maida, Narmeen Zakaria Bawany, and Kiran Fatima. "House price prediction using machine learning algorithm-the case of Karachi city, Pakistan." 2020 21st International Arab Conference on Information Technology (ACIT). IEEE, 2020.

[2] Varma, Ayush, et al. "House price prediction using machine learning and neural networks." 2018 second international conference on inventive communication and computational technologies (ICICCT). IEEE, 2018.

[3] Park, Byeonghwa, and Jae Kwon Bae. "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data." Expert systems with applications 42.6 (2015): 2928-2934.
[4] Weisberg, Sanford. Applied linear regression. Vol. 528. John Wiley & Sons, 2005.
[5] Biau, Gérard, and Erwan Scornet. "A random forest guided tour." Test 25 (2016): 197-227.
[6] Wade, Corey, and Kevin Glynn. Hands-On Gradient Boosting with XGBoost and scikit-learn: Perform accessible machine learning and extreme gradient boosting with Python. Packt Publishing Ltd, 2020.
[7] Steinwart, Ingo, and Andreas Christmann. Support vector machines. Springer Science & Business Media, 2008.
[8] Taylor, Courtney. "What Are Residuals?" ThoughtCo, Apr. 5, 2023, thoughtco.com/what-are-residuals-3126253.
[9] Chen, Tianqi, et al. “Package ‘xgboost’.” R version 90 (2019): 1-66.
[10] Biau, Gérard. “Analysis of a random forests model.” The Journal of Machine Learning Research 13.1 (2012): 1063-1095.
[11] Seber, George AF, and Alan J. Lee. Linear regression analysis. Vol. 330. John Wiley & Sons, 2003.
[12] Chang, Chih-Chung, and Chih-Jen Lin. "LIBSVM: a library for support vector machines." ACM transactions on intelligent systems and technology (TIST) 2.3 (2011): 1-27.