Research on the Prediction of Boston House Price Based on Linear Regression, Random Rorest, Xgboost and SVM Models

Yihan Chen

International School of Economics and Management, Capital University of Economics and Business, 100070 Beijing, China

* Corresponding Author Email: 1812231130@mail.sit.edu.cn

Abstract. Regression problems, which make up a significant portion of machine learning research, include the Boston housing price forecast problem. Regression problems are a subset of artificial intelligence. Regression problem-solving algorithms that are more frequently used four different machine learning models—the linear regression model, the Random Forest model based on an integrated learning technique, the XGBoost model, and the SVM model—are utilized for training and testing in this study. These models were chosen based on the real-world Boston house price prediction problem. From various angles, the effectiveness of the various algorithmic models on the regression problem of house price prediction was evaluated, and it was determined that among the many algorithmic models. For this problem, the best model is XGBoost Regression and the worst model was the SVM model. The most typical benefit of the XGBoost model is that It can capture complex non-linear relationships between variables.The disadvantage of the SVM model is its computational complexity.

Keywords: Linear Regression, Random Rorest, XGBoost and SVM Models, price forecast.

1. Introduction

Since 2011, the growth rate of real estate sales prices in Boston (real estate market) has reached more than 60% on the whole, and according to such a data, it also shows the economic development situation in Boston in recent years. The number of businesses that can be owned in Boston is gradually increasing, and the global presence of these businesses is also gradually expanding. According to such a situation, it shows the continuous good economic development of enterprises also means that the economic income of the residents who work in these enterprises is rising. Therefore, after the economic income level of residents in Boston rises, it means that they can have more economic savings to bear the rising real estate sales prices in Boston. Of course, for some consumers, they may buy real estate with higher sales price for their better living environment, which also stimulates the emergence of high-priced real estate. On the whole, the real estate sales price list in Boston is relatively high in the whole United States, and it will certainly show a higher development situation in the future. So, in this paper, different regression analysis methods will be used to study the housing price forecast in Boston, Analyzing the current development situation in Boston and the advantages and disadvantages of different forecasting methods.

From the previous literature, house price prediction is a widely investigated topic, such as, boston house price is forecasted by Tian, adopting the method of multiple machine learning algorithms [1]. By analysing the factors influencing house prices in different neighbourhoods in the Boston area in 1980, Business Forum scholars Chen and Qing used quantile regression as the basic method to initially explore the factors influencing house prices in a region in addition to the level of economic development [2]. In statistical research, data and models are two indispensable components. Based on the example of Boston house price data, statistician Jiang explains the origin of quantile regression models, The model has the benefits of both parametric and non-parametric techniques and is more versatile and adaptable than both parametric and non-parametric models in terms of both its advantages and limitations [3-4].

The reason for selecting Boston house prices is that the population of Boston has been growing for the last decade and there is a high demand for housing all year round. The reason why people

want to buy a home in Boston as an investment or to settle in Boston, even though the prices are so high, is because Boston is one of the world's economic engines and is extremely resourceful in terms of different sectors and elements. Boston has a high standard of education, with 86 institutions of higher learning in the Boston area. In addition to the world-renowned Harvard and MIT, Boston is home to Wesleyan College, Tufts College, Brandeis Landis University and Boston University, which are ranked among the best in the country. This means that talented students from all over the world come to Boston every year to study. In addition to the students who live on campus, there is of course the hassle of renting an apartment. According to research, 300,000 university students in the Boston metropolitan area need to find off-campus accommodation. In addition to this, Boston's economy is well developed and, as the traditional and largest centre for life sciences in the United States, it attracts huge amounts of venture capital funding. Boston is also the largest centre for medical research in the United States. Longwood Medical Industries is a world-renowned centre for health and medical research. This has brought a high standard of living to Boston, as well as extremely high property prices. Due to the characteristics of Boston's small metropolitan area, fast-growing resident population and large transient population, Boston has a long-term housing vacancy rate of less than 5%, which leads to a high potential for housing investment in Boston. In order to better assist people with these needs, researching Boston house prices is certainly a very practical issue that can lead to very practical solutions and references for people. Also, homeowners, investors, and real estate experts are all very interested in the Boston housing market. Predicting housing prices accurately is essential for making informed decisions and maximizing returns. In this study, various regression models will be used to analysis and forecast Boston housing prices.The prediction will be made using the training of popular algorithm models for regression issues in the prediction model utilized in this paper, and the prediction effects of linear regression, random forest regressor, XGBoost regressor and SVM regressor methods will be compared, in order to fulfill the goal of locating the best model to address this issue.

2. Data and method

2.1. Data

2.1.1 Sample and variables

To begin, The data used in this article comes from the kaggle website. The data covers home prices in Boston up to four years ago. 13 factors in Table 1 were identified as independent variables that affect Boston house prices and denoted them by their acronyms. The Boston house price data was imported into pandas and generated the following dataframe for the rest of the analysis.

Table 1. Variables and definitions

Variables	Definitions
CRIM	per capital crime rate by town
ZN	Proportion of residential land zoned for lots over 25,000 sq.ft
INDUS	Proportion of non-retail business acres per town
CHAS	Charles River dummy variable (=1 if tract bounds river; 0 otherwise)
NOX	nitric oxides concentration (parts per 10 million)
RM	Average number of rooms per dwelling
AGE	Proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centres
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per 10,000ued
PTRATIO	pupil-teacher ratio by town
B	1000(Bk-0.63) ^2 where Bk is the proportion of blacks by town
LSTAT	% lower status of the population

2.1.2 correlation analysis

Zhu suggested that the high rate of fit of historical data and the decline in forecast performance due to the instability of correlation coefficients is an urgent problem in statistical forecasting [5]. In different fields, correlation analysis has a history of different advances in multiple dimensions. The physicist Sun Tsang proposed the singular spectrum analysis method and applied it to the cross-correlation analysis of time series. The singular spectrum analysis method is first used to remove the exponential trend in the time series, and then the detrended cross-correlation analysis is performed. The simulation results show that the method can not only effectively remove the exponential trend and obtain accurate correlation scale index analysis results, but also the unique parameter window length in the algorithm has a certain adaptive range [6]. In correlation analysis of time series, Yu suggests that the estimates may have large variance and may be highly correlated with each other. Therefore, it is not possible to expect the estimates of the correlation function to be very close to the theoretical values and it is usually possible to get a fairly good idea of the broad characteristics, while the more refined features may not be representative of the actual results [7]. In order to establish an effective prediction model, the proposed method was validated by simulations using artificial mathematical models and actual time series models. The simulation results show that the proposed method can effectively perform mutual information estimation and variable correlation analysis to select the relevant input variables and finally establish a prediction model with high accuracy [8]. Liu, who has also studied house prices, has shown that house price to income ratios between different cities are correlated and show different patterns, reflecting from one side the uneven development within this region [9]. In this paper, correlation shows the interaction between time series, and it has been found that time series output from complex systems in real life are generally non-stationary and need to be de-trended before correlation analysis can be performed. Therefore, the investigation of time series de-trending correlation methods can provide insight into the underlying mechanisms of complex systems, investigate the interactions between different systems and predict changes in the system. As can be seen from figure1, correlation coefficient analysis was used to produce the following heatmap.

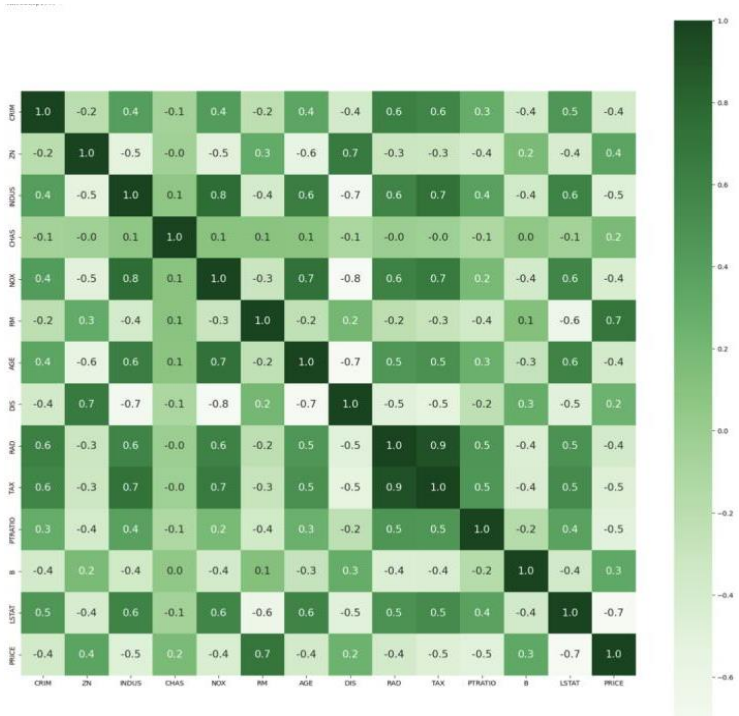

Figure 1. Heatmap

As shown in figure 1, the independent variable RM and the dependent variable PRICE have a substantial positive linear correlation coefficient of 0.7, whereas the independent factor ISTAT and the dependent variable PRICE have a stronger negative linear correlation coefficient of -0.7.

In other words, it can be deduced that the average housing price increases linearly with the average number of rooms per residence, and the greater the proportion of low-status population in the area, the lower the average housing price.

2.2. Method

To solve the problem of predicting house prices in Boston, four methods were used to build the model. They are Linear regression, Random Forest regressor, XGBoost regressor and SVM regressor. In the context of forecasting Boston house prices. Mean squared error (MSE), mean absolute error (MAE), and R-squared (R2) scores are just a few of the multifaceted, multi-perspective evaluation metrics that will be used to evaluate each model's performance and predictive ability. In addition, the paper will discuss the characteristics of each model that are helpful in this approach and the drawbacks and reasons for the poor experimental results and make recommendations for their applicability in real-world scenarios. Five steps were used to complete each model, which are data analysis, data preprocessing, separation of training set and verification set, building model and fitting model.

2.2.1 linear regression

In order to establish a linear relationship between the input elements and the target variable, one of the most well-known and often used regression techniques is linear regression. This method is also straightforward and effective. It makes the assumption that the independent and target variables have a linear connection and looks for the line of best fit to reduce the discrepancy between the expected and actual values. It assumes a linear equation of the form.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{1}$$

where Y is the dependent variable, X1, X2, ..., Xn is the independent variable, β0, β1, β2, ..., βn are the coefficients to be estimated. It does this by estimating coefficients for each input element. These coefficients represent the effect of each element on the target variable. Linear regression is known for its interpretability and simplicity, as it provides coefficients indicating the effect of each characteristic on the target variable. Using the Boston home price dataset, use linear regression and evaluate its performance in terms of predictability and interpretability. In the context of predicting Boston home prices, linear regression can be used to better understand the relationship between factors like crime rate, typical room size, and proximity to employment centers and how they affect home prices. By analysing the coefficient values, I can determine which characteristics have a significant impact on predicted prices.

Discussing the mathematical implications of regression in terms of a unitary linear equation it can be found that if there is k observations for (Xi, Yi) and want the equation to coincide as closely as possible with the observed data.

Suppose the fitted equation is y = a*x + b. Substituting x from the observed data gives y1 = a*x + b and the error in y from the observed data is error = y - y1 = y - (a*x + b).

Let the objective function be.

$$c = \sum_{i=1}^{k} error_i^2 = \sum_{i=1}^{k} [y_i - (a * x_i + b)]^2 \tag{2}$$

In order to make error smaller, i.e., when C takes a very small value, the bias derivative for a and b is 0:

$$\begin{cases} \frac{\partial c}{\partial a} = -2 * \sum_{i=1}^{k} (y_i - a * x_i - b) * x_i = 0 \\ \frac{\partial c}{\partial b} = -2 * \sum_{i=1}^{k} (y_i - a * x_i - b) = 0 \end{cases} \tag{3}$$

Solve for:

$$\begin{cases} a = \frac{\sum_{i=1}^{k} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{k} (x_i - \bar{x})^2} \\ b = \bar{y} - a * \bar{x} \end{cases} \tag{4}$$

This gives the regression equation:

$$y = a * x + b \quad (5)$$

Further, it can be also derived a multiple regression equation using multiple independent variables.

In the application of predicting Boston house prices, I first calculated the y-intercept value, which is 36.357, and then converted the coefficient value to a data frame.

The advantages of linear regression are threefold. The first point is simplicity. Linear regression is easy to understand and interpret. The coefficients shed light on the strength and slant of the investigated relationship between the independent and dependent variables. The second benefit is that it moves quickly. Because linear regression is computationally effective, it can handle massive data sets. The ability to interpret is the third point. Characteristic importance analysis is made possible by the model's provision of characteristic coefficients.

Because it presupposes a linear relationship between variables, linear regression has the drawback that it may not hold true in complex real-world circumstances. Furthermore, linear regression has a low level of complexity. The non-linear correlations between variables may be difficult for linear regression to capture. Finally, outliers might affect linear regression results. The model's performance might be impacted by this.

2.2.2 Random forest regressor

The Random Forest Conditioner is a hybrid learning approach that brings together the benefits of several models and has the ability to mix different decision trees to produce predictions. The data used to train each decision tree is chosen at random. Its final prediction technique is created by averaging all of the trees' forecasts. Because they can successfully handle non-linear relationships and feature interactions, random forest moderators are becoming more and more common today. The model can handle both numerical and categorical features and performs well in capturing complex non-linear relationships and handling outliers. It is a powerful algorithm for capturing complex relationships and interactions between various features in the case of house price prediction with the Random Forest regressor. This integrated approach helps reduce overfitting and improves generalisation. In addition, the random forest regressor is robust to outliers and missing values. It can provide insights into the importance of features, indicating which variables have the greatest impact on house prices.

In the case of random forests, decision trees are used as the base model. The decision tree splits the data based on feature thresholds and makes predictions based on the majority vote of the leaf nodes. Random forests are robust to over-fitting, handle large numbers of features well and capture complex relationships in the data. However, they are computationally expensive and may not provide models that can be easily interpreted. The performance of random forest models can be assessed using metrics such as MSE, MAE and R2. In addition, the importance of features can be analysed to understand the relative importance of different features in predicting house prices. The formula for predicting the output of a random forest can be expressed as

$$y = f_1(x) + f_2(x) + \cdots + F_m(x) \quad (6)$$

Where y is the predicted output. f1(X), F2(X), ..., Fm(X) are the individual predicted values for each decision tree in the random forest.

For the problem of predicting Boston house prices, the values of R^2 can also be used, Adjusted R^2, MAE, MSE and RMSE to evaluate the stochastic forest regressor model. The results of the comparison are shown in the table below.

The advantages of the random forest regressor are likewise threefold. The first is flexibility. Random forests can handle a large number of features and capture non-linear relationships efficiently. The second is robustness. They are less prone to over-fitting, as they average the predictions of multiple trees. The third is feature importance. Random forests provide feature importance rankings, allowing insight into which features contribute most to the prediction.

There are also three drawbacks to the random forest regressor. The first is interpretability. Compared to linear regression, random forests lack interpretability because they involve an ensemble of decision trees. The second is computational complexity. The computational cost of training random forests can be high, especially for large numbers of trees or features. The third is memory consumption. Random forests can consume large amounts of memory resources, especially for large datasets.

2.2.3 XGBoost regressor

The XGBoost regressor is an optimised gradient boosting algorithm that has become popular in machine learning competitions due to its superior performance. It is an integrated learning technique that builds an ensemble of weak decision tree models in a sequential manner and has proven to be very effective in a variety of regression tasks. The application of XGBoost to the Boston house price dataset and evaluate its predictive performance compared to other models. The impact of hyperparameter tuning on model accuracy have been discussed. XGBoost works by initially fitting a single decision tree to the data and then sequentially adding more trees to correct for errors made by previous models. It uses gradient descent optimisation to minimise the loss function and improve the predictions of the model. xGBoost is known for its efficiency, scalability and ability to handle complex functional interactions. It also incorporates regularisation techniques to prevent over-fitting.

The main features and benefits of XGBoost are mainly the advantages offered, such as handling missing values, regularisation techniques to prevent over-fitting, and support for parallel processing. It also provides built-in evaluation metrics and feature importance analysis. In terms of parameter tuning and model optimisation, XGBoost involves tuning hyperparameters such as learning rate, tree depth and the number of estimators to optimise model performance. Techniques like cross-validation and grid search can be used to find optimal parameter values.

The formula of XGBoost regressor uses gradient boosting, and the prediction formula can be represented as:

$$Y = \sum w_i * f_i(x) \tag{7}$$

Where Y is the predicted output. wi is the weight assigned to each weak learner. fi(X) is the prediction from each weak learner.

Gradient boosting, which sequentially combines weak learners to produce a powerful predictive model, is enhanced in XGBoost. A new weak learner is fitted to the residuals (errors) of the preceding model at each step via the procedure. The total of all the weak learners' guesses, weighted by importance, forms the final forecast. The advantages of the XGBoost regressor are fourfold. The first is high performance. XGBoost is known for its scalability and efficiency, making it suitable for large data sets. The second point is the handling of complex relationships. It can capture complex non-linear relationships between variables. The third point is regularisation. XGBoost includes regularisation techniques to prevent overfitting and improve generalisation. The fourth point is the built-in evaluation metric. Evaluation metrics such as Mean Square Error (MSE) and Root Mean Square Error (RMSE) are provided during training.

It has two drawbacks. The first is the computational resources. XGBoost requires more computational resources than simpler models such as linear regression. The second is the complexity of tuning. Tuning the hyperparameters of XGBoost requires careful optimisation and can be very time consuming.

2.2.4 SVM regressor

A potent method utilized mostly for classification tasks, the Support Vector Machine (SVM) regressor can also be used to solve regression issues. It is also applicable to regression issues, where the SVM regressor translates the data to a high-dimensional feature space with the goal of locating the optimum hyperplane that maximizes the distance between data points. Being used in the context of Boston house price forecasting, the SVM regressor can identify the best hyperplane that best fits the data and minimises error. It is particularly effective in dealing with high-dimensional data,

allowing linear and non-linear relationships to be captured using kernel functions. the SVM regressor can provide valuable insights into non-linear relationships in housing data and may be effective in identifying outliers.

The data was first pre-processed by dealing with missing values and normalized attributes, and then it was analyzed to determine the Boston house price forecast. A training set and a test set were created from the dataset. The training set will be used to train each of the four models (linear regression, random forest regressor, XGBoost regressor, and SVM regressor), which will then be tested using the test set and the relevant evaluation metrics (such as mean squared error (MSE) or R-squared score).

The advantages of the Support Vector Machine regressor are threefold. The first is versatility. SVMs can use different kernel functions to deal with linear and non-linear relationships. The second point is robustness to overfitting. SVMs use regularisation parameters to prevent overfitting and to handle noisy data. The third point is its effectiveness in high-dimensional spaces. sVM performs well in high-dimensional spaces, making it suitable for datasets with many features. The downside is the computational complexity; SVMs can be computationally expensive, especially for large datasets. In addition, there is the choice of kernel. Choosing the right kernel function and tuning the relevant parameters can be challenging. And interpretability. In contrast to linear regression, SVM models may not provide intuitive explanations.

2.2.5 Prediction performance indicator

To assess the model, five values were computed. The linearity of the relationship between X and Y is measured by R2. It is also known as the percentage of the dependent variable's volatility that can be predicted using the collection of independent factors. For regression models with various numbers of predictors, the adjusted R2 is the adjusted R-squared, which provides comparative explanatory power. The mean of the absolute values of the errors is known as the MAE. It measures the difference between two continuous variables, in this case the actual and predicted values of y. MSE is the mean squared error (MSE) as with MAE but requires the difference to be squared before they can be summed, rather than using absolute values. rMSE is the mean squared error (MSE) as with MAE but requires the difference to be squared before they can be summed, rather than using absolute values. In a technical note, T. Chai demonstrates that the meaning of RMSE has been explained more precisely. When a Gaussian error distribution is anticipated, RMSE provides a more accurate picture of model performance than MAE. They also demonstrate that the RMSE meets the distance metric's condition for the triangle inequality. They do not believe that the MAE is less effective than the RMSE. Instead, it is frequently necessary to use a variety of measures, including but not limited to RMSE and MAE, to evaluate model performance [10].

3. Result

3.1. Prediction results of Linear regression

It can be seen that the model's evaluation scores are practically identical to the train data when comparing the model's projected train data with the test data (table 2). The model is therefore not overfitting.

Table 2. Results for LR

	R^2	AdjustedR^2	MAE	MSE	RMSE
Train data	0.747	0.737	3.090	19.074	4.367
Test data	0.712	0.685	3.859	30.054	5.482

Then the differences were visualized between actual prices and predicted values (figure 2) and checked residuals (figure 3) as the graphs below:

Figure 2. Prices vs predicted prices.

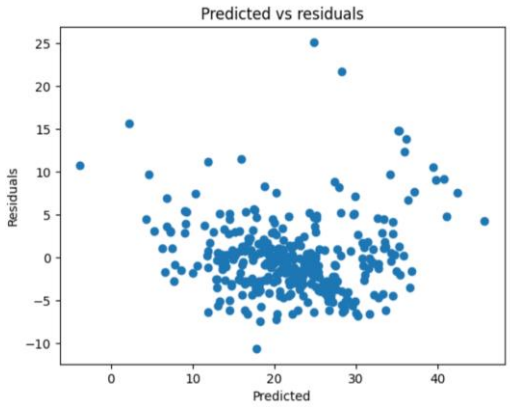

Figure 3. Predicted vs residuals.

3.2. Prediction results of random forest regressor

For the problem of predicting Boston house prices, the values of R ², Adjusted R ², MAE, MSE and RMSE are also be used in this paper to evaluate the stochastic forest regressor model. The results of the comparison are shown in the table 3.

Table 3. Results for RFR

	R^2	AdjustedR^2	MAE	MSE	RMSE
Train data	0.982	0.981	0.814	1.348	1.161
Test data	0.820	0.803	2.550	18.825	4.339

Figure 4 and figure 5 were the visualising difference between the actual price and the predicted value and checked the residual values.

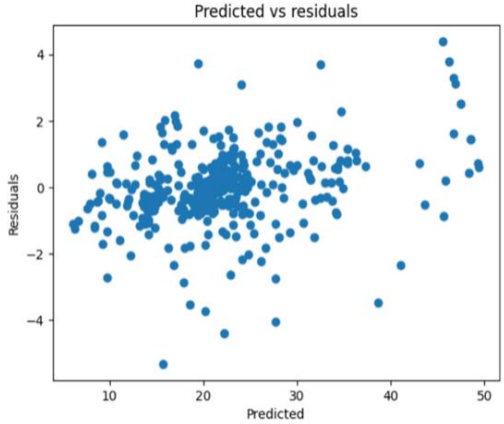

Figure 4. Predicted vs residuals

Figure 5. Prices vs predicted prices.

3.3. Prediction results of XGboost regressor

The results are shown in Table 4.

Table 4. Results for XGBoost

	R^2	AdjustedR^2	MAE	MSE	RMSE
Train data	0.642	0.628	2.936	26.954	5.192
Test data	0.590	0.551	3.756	42.811	6.543

Figure 6 and figure 7 were the visualising difference between the actual price and the predicted value and checked the residual values.

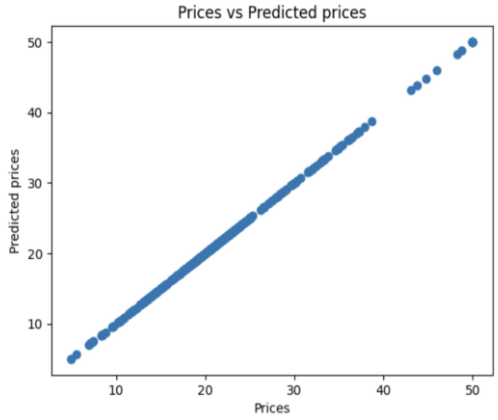

Figure 6. Prices vs predicted prices

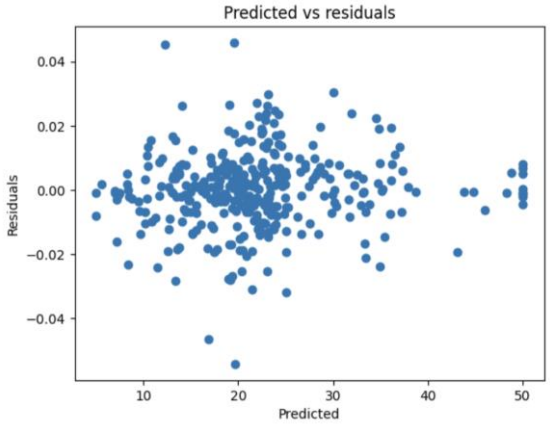

Figure 7. Predicted vs residuals.

3.4. Prediction results of SVM regressor

The results are shown in Table 5.

Table 5. Results for SVMR

	R^2	AdjustedR^2	MAE	MSE	RMSE
Train data	0.999	0.999	0.009	0.0001	0.012
Test data	0.858	0.845	2.530	14.828	3.851

Figure 8 and figure 9 were the visualising difference between the actual price and the predicted value and checked the residual values.

Figure 8. Prices vs predicted prices

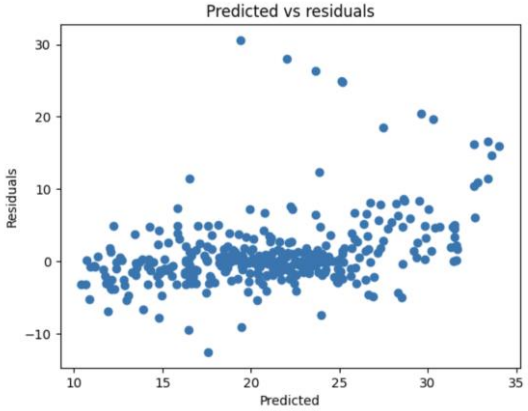

Figure 9. Predicted vs residuals.

3.5. Comparison of prediction performance of 4 methods

Finally, the R-squared Score was used to indicate the fitness of each model. A higher score means that the method is more applicable to solving the problem has been studied. As shown in table 6, XGBoost Regression works the best for this dataset.

Table 6. R-squared score

	Model	R-squared Score
2	XGBoost	85.799520
1	Random Forest	81.971735
0	Linear Regression	71.218184
3	Support Vector Machines	59.001585

4. Conclusion

In this paper, the Boston house price problem was studied using linear regression models, Random Forest models, XGBoost models and SVM models. The main conclusions obtained were firstly that the XGBoost Regression method was the best choice, and the worst was the Support Vector Machines Secondly, the advantages of the XGBoost model in terms of high performance, scalability, efficiency, handling of complex relationships and regularisation techniques make it suitable for large data sets, capturing complex non-linear relationships between variables, preventing over-fitting and improving generalisation. The shortcoming of the SVM model is its inability to provide intuitive explanations. These features may have contributed to this result.

The research in this paper provides help and methodological reference for those who want to buy a home and need to predict the price of a home in Boston. The limitation is that this paper does not take into account factors such as educational resources and the level of greenery in the surrounding environment in its analysis of predicting house prices in Boston, and subsequent studies can continue to refine this point.

References

[1] R. Z. Tian. Internet Health, **21**, 228-230, (2019)

[2] Z. L. Chen, Q. Qing. Business Forum (Industrial Economics), **30**, (2015)

[3] Y. J. Jiang. Southwest University of Finance and Economics, **712**, (2015)

[4] Y. M. Weng. Master's Thesis (Econometrics), Xiamen University, (2009)

[5] S. M. Zhu. Acta Meteorologica Sinica, 4**0**,4(1982)

[6] Z. G. Sun, S. Cheng, F. Wang. Journal of Liaoning Normal University (Natural Science Edition), **41**, 2(2018)

[7] N. L. Yu, Y. Yi, X. Q. Tu, Mathematical Theory and D. Applications, **27**, 1(2007)

[8] Z. P. Liang. Dalian University of Technology Master's Thesis (Control Theory and Control Engineering), **61**, (2011)

[9] S. K. Liu, Industry Perspectives (Business Edition), **51**, (2021)

[10] T. Chai, R. R. Draxler, Geoscientific Model Development, **7**, 3(2014).