Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

A Comparative Study of Regression Models for Housing Price Prediction

Zizheng Li*

School of Mathematics Physics and Statistics, Shanghai Polytechnic University, Shanghai, China

*Corresponding author: 20211120138@stu.sspu.edu.cn

Abstract. The property market is closely related to the regional economy. This study focuses on exploring the problem of house price prediction in the property market. This paper aims to reveal the performance of extreme gradient boosted tree regression, ridge regression, decision tree, and random forest regression models in feature selection and data processing, and assess their effectiveness in house price prediction. Aspects compared include prediction accuracy, data required for fitting, and running time. The study shows that the random forest and decision tree regression models perform best. Although these two models require a certain amount of data, they are easily satisfied in the context of big data. Although the fit of the random forest regression model and the decision tree regression model is similar, the training time required for the random forest regression model is too long. Therefore, it is concluded that the decision tree regression model performs well on general real estate datasets and is a widely applicable regression model in the real estate domain. This study can provide a helpful reference for data analysis and forecasting in the real estate field and provide a basis and inspiration for further research in related fields.

Keywords: Regression; Machine Learning; Housing Price Prediction.

1. Introduction

House prices are often seen as essential for assessing economic development and market trends [1]. When the economy is booming, people are more able to buy property and house prices will trend upwards; conversely, house prices will trend downwards. Property is a major form of investment for many people, and fluctuations in house prices can reflect the overall supply and demand in the property market, investor confidence and changes in the financial markets. For individuals and families, property is usually one of the largest assets. Therefore, fluctuations in house prices also affect the financial situation of many individuals and households, affecting society's consumption and investment behaviour.

This impact makes the assessment and prediction of house prices not only an economic and market issue, but also involves aspects of urban planning and resource allocation. House prices vary in different areas of the city, and the level of house prices affects the city's social structure, the mobility of residents and the allocation of regional resources. In areas with high house prices, the urban development process tends to be more active, so more investment and resources will enter the area, leading to a further rise in the cost of living and affecting the social structure and resource allocation. At the same time, changes in house prices have a profound impact on the migration of urban residents, with lower house prices often attracting investment and leading to population inflows. In comparison, higher house prices inhibit the migration of foreign populations and weaken the city's ability to attract foreign investment.

Assessing and forecasting house prices is therefore of great relevance. Detecting and analysing house prices helps the economic system and society to evaluate and correct themselves. For society, the prediction of house price can infer the market trend, increase the economic gain, weaken the damage of market fluctuation to its own economy, and thus find the direction of development; for the government, the monitoring of house price can understand the regional economic situation, adjust the policies promptly, such as the real estate tax policy, the housing subsidy policy, and so on, to avoid the social stability problems caused by the change of house price, and to maintain the social fairness

and stability. Urban development departments can use housing prices to regulate urban development gaps, control resource tilting, and ensure balanced regional development; for individuals, assessing and predicting housing prices can help optimize investment strategies and financial planning and reduce investment losses.

As the changes in house prices are affected by many aspects, such as government policies, supply and demand, employment levels, etc., the traditional manual extrapolation often fails to get accurate results. With the development of computer technology, machine learning algorithms are gradually utilised to solve the problem of house price prediction. Machine learning algorithms can handle large amounts of complex data and build house price prediction models through statistical analysis and pattern recognition. Supervised learning algorithms such as linear regression and decision trees can predict future house price trends based on historical data [2]. Unsupervised learning algorithms such as cluster analysis and dimensionality reduction algorithms can help to find underlying market patterns.

Nor Hamizah Zulkifley et al [3] trained different models using house price datasets, and the results indicated that artificial neural networks, support vector regression, and other models were superior.Satish, G. Naga et al [4], considered that house price predictions obtained by lasso regression were superior, considering the accuracy condition. Meanwhile, research by Y. Zhou selects XGBoost as the best model since it provides the lowest RMSE value in contrast with other models in his study [5]. However, there are often large differences between different house price datasets due to different data sources, and differences in input features and dimensions can affect the performance of machine learning algorithms.

Based on the above background, this paper will aim to discuss and analyse the performance of different machine learning algorithms on the house price prediction problem. Firstly, this paper takes the US real estate dataset as the research object and selects different linear regression models and tree models for comparative analysis. Then, the model accuracy is evaluated in terms of mean square error, mean absolute error, etc. Meanwhile, its learning ability and calculation time are compared with control variables and presented graphically. Finally, the experimental results are analysed and summarized to find a better solution algorithm for the problem under general conditions.

2. Data and Methods

2.1. Data Source

The data of this paper comes from News Corp subsidiary Move, Inc, which is about the basic indicators of real estate transactions, including property area, transaction amount and so on. The wide range of data sources can better reflect the entire U.S. real estate economy and is representative. Using this dataset as the data source for testing each model can better reflect the algorithm's performance on the general dataset. In order to deal with the noise in the dataset and reduce the interference of outliers with the model's ability, data within two standard deviations of the mean in the numerical columns are retained and the nulls are filled in as column means.

2.2. Models

The following four models were selected for comparison in this paper:

a. Extreme Gradient Boosting Tree Regression Model

Gradient boosting was created by [6] in 1999 and is a commonly used machine learning algorithm. XGBoost (eXtreme Gradient Boosting) is an efficient integrated learning algorithm that combines Gradient Boosting and regularisation techniques for solving regression and classification problems.XGBoost progressively improves the model's performance by iteratively training multiple decision trees. Each tree attempts to capture the residuals (or gradients) of samples that were not correctly predicted by the previous tree, thus continuously improving the model's predictive power.

Also XGBoost uses a greedy algorithm to construct the trees. The optimal splitting point is chosen during the construction of each tree to minimise the loss function. Due to the complexity of training, XGBoost requires a lot of computational resources and time, especially when dealing with large amounts of data. Also when dealing with data that is too small or too complex, XGBoost can suffer from overfitting, leading to its poor performance on the test set.

b. Ridge Regression Model

Ridge Regression is an extension of linear regression, which solves the problem of multicollinearity in Ordinary Least Squares (OLS) by introducing the L2 regularisation term, the goal of Ridge Regression is to minimise the loss function, which is as follows:

$$L(w) = \frac{1}{2m}\left((y - Xw)^T(y - Xw)\right) + \lambda w^T w \tag{1}$$

The loss function consists of a data-fitting term and a regularisation term. The former measures how well the model fits the data, and the latter controls the model's complexity and prevents overfitting. However, this also means that ridge regression does not automatically perform feature selection; even if certain features are uncorrelated with the target variable, ridge regression still retains them and shrinks them, which may lead to an overly complex model. Since the loss function of ridge regression is based on the squared error, the performance of ridge regression decreases dramatically when the number of features is much larger than the number of samples or when there are outliers.

c. Decision Tree Regression Model

Decision tree regression model is a regression method based on decision tree algorithm that divides the input space into regions and fits a simple linear model on each region to predict the value of the target variable. The division is determined by a threshold of features to minimise the mean square error of the samples within each region. Decision tree regression models can handle non-linear relationships as well as complex data structures well because they automatically handle interactions and non-linear relationships between features. However, the decision tree model is very sensitive to noise and outliers and is also prone to overfitting problems.

d. Random Forest Regression Model

Random forest regression model is a model based on the regression forest algorithm, which randomly selects samples and features to construct independent decision trees during the training process, and the selected samples and features determine each decision tree. In the prediction phase, the random forest model averages the predictions of each tree as the final result. This creates a better generalisation ability and robustness of the regression forest model, making the random forest model less prone to overfitting. However, simultaneously, the complexity of the model also makes the random forest model training more consuming in terms of computational resources, with a long training time and reduced interpretability.

2.3. Evaluation Indicators

Based on the type of problem and model selection, the following four metrics are chosen to evaluate the model (n is the number of samples, y_i is the actual value, y_i is the predicted value, and y_i is the average of the true values.) :

a. Mean Absolute Error (MAE):

The average of the absolute errors between the predicted and true values. The formula is as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_1| \tag{2}$$

The smaller the MAE, the smaller the prediction error of the model.

b. Mean Square Error (MSE):

The average of the squares of the errors between the predicted and true values. The formula is given below:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_1)^2 \tag{3}$$

The smaller the MSE, the smaller the prediction error of the model.

c. Root Mean Square Error (RMSE).

The square root of the MSE is the prediction error's standard deviation. The formula is as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_1)^2} \tag{4}$$

The smaller the RMSE, the higher the predictive accuracy of the model.

d. Correlation Coefficient (R^2):

Also known as the coefficient of determination, indicates the correlation between the predicted and true values of the model. The formula is given below:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5}$$

The closer R^2 is to 1, the better the model fits; the closer R^2 is to 0, the worse the model fits.

3. Analysis of results

When the ratio of training set data to test set data is 4:1, the mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE), and correlation coefficient (R^2) of the four models are shown in Fig. 1. From Fig. 1, it can be concluded that the random forest regression model and the decision tree regression model are much better than the ridge regression model and the extreme gradient boosted tree regression model in terms of mean absolute error (MAE), mean square error (MSE) and root mean square error (RMSE) in the present dataset. With the correlation coefficient (R^2) as a criterion, the difference between the two decision tree regression models and the random forest regression model is not much at 0.8, slightly better than the extreme gradient boosted tree regression model at 0.5, and the ridge regression is the worst at less than 0.2. This suggests that the ridge regression does not fit the house price dataset well, and the tree models, such as the decision tree and the random forest, have good fitting results. This may be because the linear relationship between house price data features is weak, while the complex non-linear relationship is strong. Therefore, the decision tree regression model and the random forest regression model are optimal when mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE), and correlation coefficient (R^2) are used as the criteria.

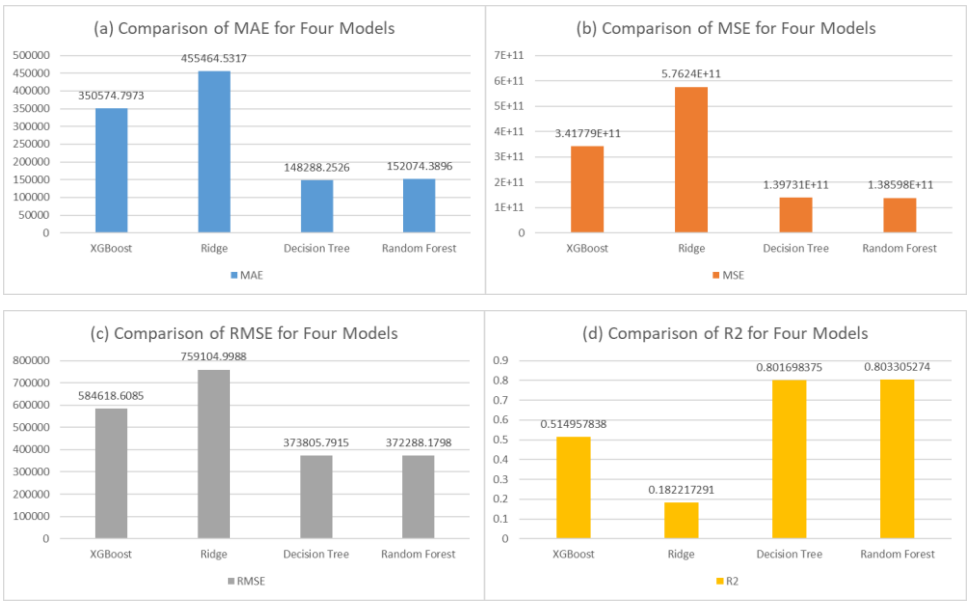

Fig.1 Comparison of MAE, MSE, RMSE, and R2 for Four Models

The mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE), and correlation coefficient (R^2) of the four models are shown in Fig. 2 as the number of data in the test set rises. As can be seen from Fig. 2, the fit of the decision tree regression model and the random forest regression model gradually decreases with the increase in data in the test set. In contrast, the fit of the ridge regression model and the extreme gradient boosting tree model remains basically unchanged, which indicates that the decision tree regression model and the random forest regression model need a certain amount of data in the training set to abstract the relationship between the data features to achieve a better fitting effect. The degree of fit of these two models continues to increase with more training data. On the contrary, the ridge regression model and the extreme gradient boosted tree model are less affected by the amount of data in the training set, and their fit is basically unchanged in the process of increasing the amount of data in the test set, which indicates that these two models do not need too much house price data to achieve a better fit.

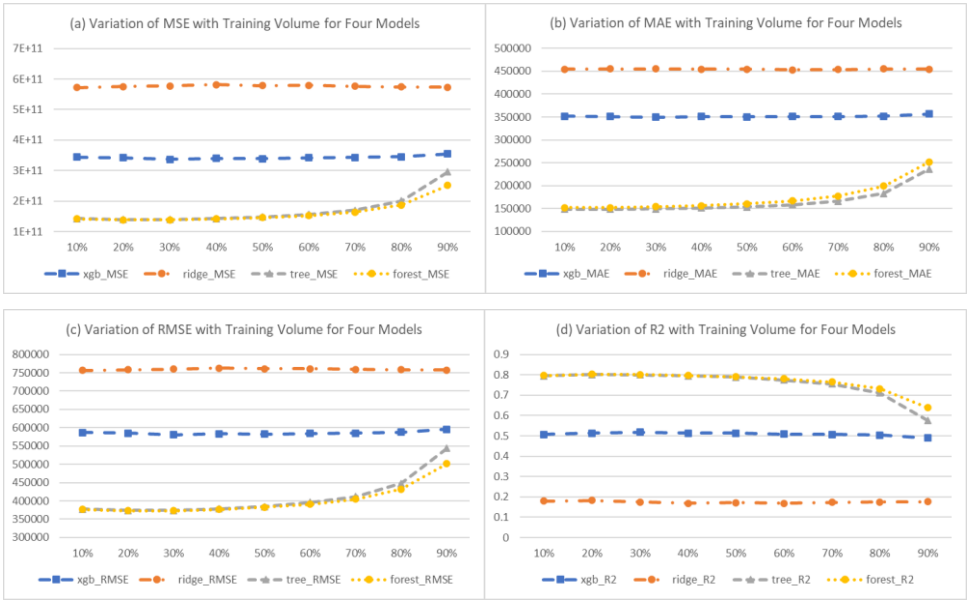

Fig.2 Variation of MAE, MSE, RMSE, and R2 with Training Volume for Four Models

As the proportion of data in the training set rises, the change in the time required from training to prediction completion for the four models is shown in Fig.3. In Fig.3, as the percentage of the number of training sets rises, the training and prediction time for the extreme gradient boosted tree regression model shows an overall increasing trend, but the increase is not significant, up to less than 1.3

seconds,which has similarities with findings of G. Naga Satish et al [7]. The ridge regression model training and prediction times vary within 0.05 seconds. Unlike the other models, the random forest regression model showed a strong upward trend and took a very long time, reaching nearly 100 times the time the other models took. The decision tree regression model also increased in time with training, reaching a maximum of 1.4 seconds.

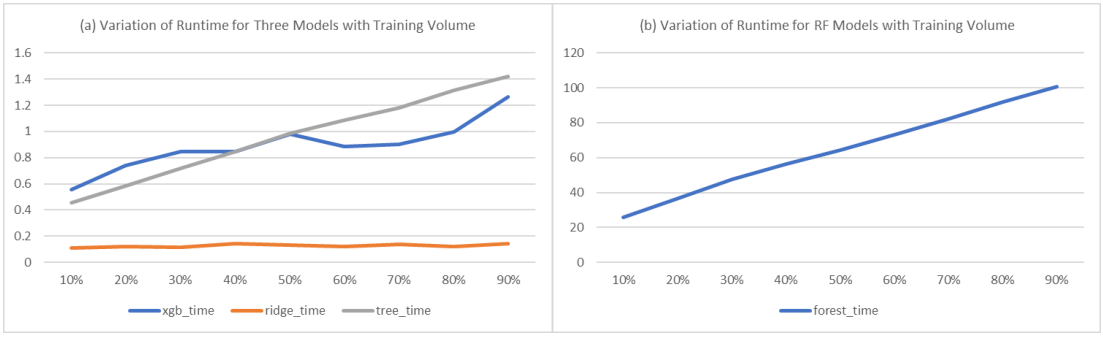

Fig.3 Variation of Runtime for Four Models with Training VolumeSummary

4. Conclusion

In this paper, four common models, Extreme Gradient Boosted Tree Regression Model, Ridge Regression Model, Decision Tree Regression Model and Random Forest Model, were selected for comparison under a general dataset. However, due to the limitations of feature selection, using only a few features for prediction may result in the model not being able to adequately capture the complexity and variability of the data, thus ignoring the model's ability to capture important features. This in turn affects the accuracy and generalisation ability of model predictions on other datasets. Therefore, subsequent experiments can add more features to fully consider the learning ability of the model as well as its prediction accuracy on complex datasets. As for the limitation of model selection, the models selected in this paper do not cover the mainstream regression models, and each model has its unique advantages and disadvantages, and only completing the comparison of the four models may ignore some models with great potential on the house price dataset. This problem can be avoided by trying more models on the general house price dataset. Noise on the dataset is dealt with in this paper, but the handling of noise is often subjective and limited, and differences in the way it is handled tend to affect the performance of the models. Also, the extent and impact of noise tends to vary depending on the dataset's characteristics, and noise often cannot be eliminated, which could potentially affect the model's fit. Therefore different noise reduction methods can be tried to observe how different noise reduction methods on different datasets affect the model. After the above comprehensive experiments, it is possible to find the model that performs optimally in the problem of house price prediction while minimising the error, which helps researchers and social workers better understand the market and promote social development.

References

[1] A. S. Temür, M. Akgün, and G. Temür. Predicting Housing Sales in Turkey Using Arima, Lstm and Hybrid Models. J. Bus. Econ. Manag., vol. 20, no. 5, pp. 920–938, 2019, doi: 10.3846/jbem.2019.10190.

[2] R. E. Febrita, A. N. Alfiyatin, H. Taufiq, and W. F. Mahmudy. Data-driven fuzzy rule extraction for housing price prediction in Malang, East Java. 2017 Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS 2017, vol. 2018-Janua, pp. 351–358, 2018, doi: 10.1109/ICACSIS.2017.8355058.

[3] Zulkifley, Nor Hamizah et al. House Price Prediction using a Machine Learning Model: A Survey of Literature. International Journal of Modern Education and Computer Science, 2020.

[4] Satish, G. Naga et al. House Price Prediction Using Machine Learning. International Journal of Innovative Technology and Exploring Engineering, 2019.

[5] Y. Zhou. Housing Sale Price Prediction Using Machine Learning Algorithms. 2020.

[6] J. H. Friedman. Stochastic Gradient Boosting. vol. 1, no. 3, pp. 1–10, 1999.

[7] G. Naga Satish, Ch. V. Raghavendran, M.D.Sugnana Rao, Ch.Srinivasulu. House Price Prediction Using Machine Learning. International Journal of Innovative Technology and Exploring Engineering, 2019.