Proceedings of the 3rd International Conference on Computing Innovation and Applied Physics
DOI: 10.54254/2753-8818/26/20241019

Machine learning on USA house price prediction

Zirong Jin

Nanyang High School, Shanghai, China

13120886825@lfcmsq.wecom.work

Abstract. Nowadays, an increasing number of students are opting to study abroad in order to acquire more advanced knowledge and pursue a superior educational environment. In many foreign countries, the option to apply for school dormitories is only available during the first year of university or graduate school. At other times, international students have to search for rented apartments or apply to stay with local host families. However, when studying abroad for an extended period, purchasing a property can potentially result in significant savings compared to renting. Therefore, this study focuses on comparing three types of machine learning techniques: multiple linear regression, Random Forest, and XGboost in predicting house prices in the United States. This research could provide reference for families studying abroad or property investors. Based on the preliminary findings of this study so far, it can be concluded that the XG-boost model demonstrates the highest accuracy and stability among these three methods.

Keywords: Linear regression model, random forest model, XGboost model, USA house price prediction.

1. Introduction

Housing price has always been a concern in society. The long-lasting pandemic named COVID-19 has already brought serious crises for global finance as well as real estate [1]. During the epidemic, real estate sales were restricted, and the number of housing transactions plummeted in most cities. Also due to the impact of the epidemic, the original rental just needs to pay rent due to the epidemic can not be put back, but still have to pay rent, yes, some people have a purchase plan. Plus the popularity of studying abroad still remains at a high level, and renting a house in a foreign country is a must. But when compared with buying a house in America, renting a house may cost 10 times more money than the former [2]. So the assessment of house prices becomes increasingly important for investors, creditors, and the government to pay attention to. Multiple linear regression, as a statistical analysis method in mathematical statistics, has been used in the prediction of housing prices because of its three advantages: fast modeling speed, the explanation of each variable can be given according to the coefficient, and the sensitivity to outliers. In order to fit the data more effectively, on the basis of the original data, some variables such as the current unemployment rate, national consumption index, etc., are added. The success of the model and the subsequent extensive reference value can be reflected by the experiment that the residual error predicted by the model is about 2% [3]. There are also studies that use multivariate linear regression to qualitatively analyze each factor and confirm the wireless relationship between them, and then establish BP neural network analysis to predict housing prices. However, in the process, the problem of long training time and low prediction accuracy is encountered because the neural network

DOI: 10.54254/2753-8818/26/20241019

cannot eliminate the redundant connections among many influencing factors. Finally, the researchers used the PCA algorithm to pre-select variables with relatively strong features to solve the problem[4]. However, this method makes the research process more complicated. Therefore, the author collected the US housing price data set from the kaggle website in May 2014 as the training set and test set of this study, and used three models, multiple linear regression, random forest, and XGboost, to train the model. And compared these models, trying to find relatively simple, efficient, high prediction accuracy of the model, and the defects of each model after training to speculate and analyze.

2. Data description

The data is searched on the Kaggle website (https://www.kaggle.com/datasets/shree1992/house data.). This data has 21613 rows and 14 columns in total. Because of the large and complex data, machine learning is used here to process it efficiently and accurately. It lists home prices by region in the United States for the period up to May 2014. The price column will be considered as the dependent variable, and the remaining 13 columns will be considered as independent variables. Fortunately, all the data is complete and there are no null values, so the normal calculation does not need to be filled or omitted.

The following diagram Figure 1 is a thermodynamic relationship between multiple independent variables and one dependent variable. There are 12 independent variables in total: bedrooms,bathrooms,sqft_living,floors,waterfront,view,condition,grade,sqft_above,sqft_basement,lat, sqft_living15 and sqft_lot15.The dependent one is house price.The authors show the closeness of the relationship between the variables as the shade of blue. It can also be seen from this chart that the price is greatly affected by the number of bathrooms, the living area in the house, the number of house grades, and the area of the living room and basement. This figure also shows the relationship between various independent variables, which will serve as a reference for future data exploration.

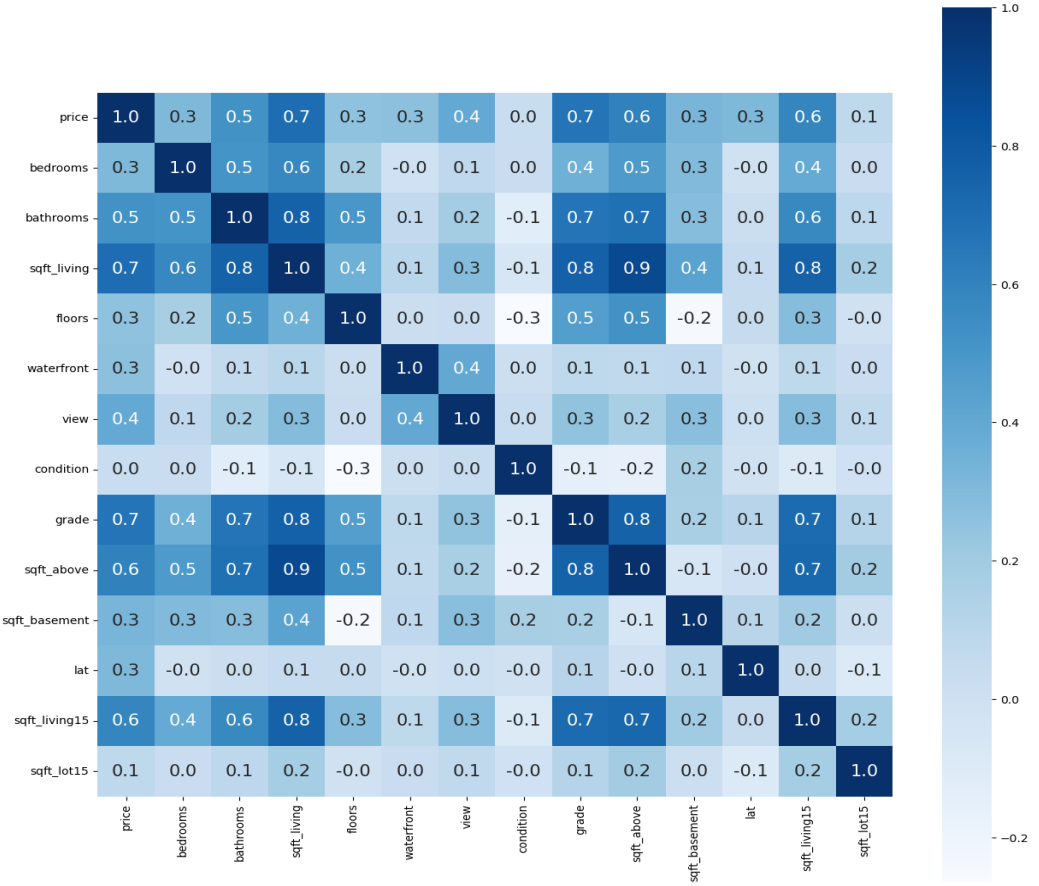

Figure 1. Thermodynamic relationship between the variables

DOI: 10.54254/2753-8818/26/20241019

3. USA House Price Prediction

3.1. Multiple Linear Regression

Multiple Linear Regression is often used in situations when various variables are all contributed to one variable and these independent variables are mutually exclusive with each other [5]. These independent variables are valid ones so each of them is put into processing.

The formula of Linear Regression is y=-271438.01x1-3558.54x2+122.63x3-15703.38x4+557519.68x5+63942.34x6+55033.58x7+83423.39x8+68.11x9+54.52x10+673949.63x11+8.37x12-0.4x15-32640413.5.The x stands for all independent variables mentioned above and each coefficient is corresponding with them which sequence are also listed above.The weights of each are presented in figure2 y stands for price.

3.1.1. Process. The whole data are separated into train data and test data, which contain 0.8 and 0.2 in random order respectively. The model needs to be tested by statistical methods such as linear relationship of equations, significance of coefficients, and distribution of residuals before it can be used to analyze and explain practical problems. r square is a statistic that measures the degree of fit of multiple regression equations. The r^2 is closer to 1, the better the proportion of the linear part of the dependent variable can be explained by the linear part of the independent variable. In order not to seriously affect the dependent variable when adding more independent variables, it is better to use adjusted r^2 [6]. MAE measures the mean of the absolute error between the predicted value and the actual value.(MAE=1/n∑|Yi-Xi|).MSE refers to the expectation of the square of the difference between the parameter estimate and the parameter value. RMSE is an index used to measure the prediction accuracy of a prediction model on continuous data. It measures the root-mean-square error between the predicted value and the true value.

Figure 2. Linear relationship between predicted house prices and real house prices

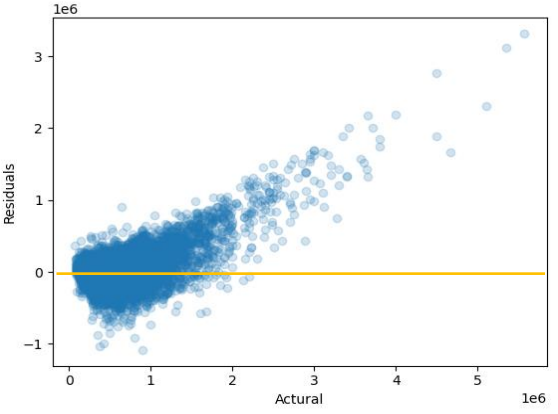

Figure 3. Relationship between true value and residual

3.1.2. Result Analysis. The first scatter plot shows the linear relationship between the actual price and the predicted price. It is clear that according to Figure 2, it doesn't exhibit an evident linear shape. Still, r^2 doesn't present a high level, which is just 0.6711 just so the degree of fit is not good.In order to reduce the error caused by outliers, the authors use adjust r^2:0.6708. And MAE,MSE,RMSE are 130410.5426,41265086241.8517,203138.0965 respectively.As is shown in Figure 3, the scatter plot indicates the relationship between the actual price and residuals. As the actual price becomes higher, residuals keep getting farther and farther away from 0, which means a more exaggerated error it has. The Multiple Linear Regression model is easily affected by individual abnormal data. The problem that the model does not meet its practical significance often occurs in the case of abnormal data, resulting in an unreasonable relationship coefficient between the independent variables [7].

DOI: 10.54254/2753-8818/26/20241019

3.2. Random Forest Regression

A fundamental technique for classification and regression, decision trees display rules such as what values are acquired under what circumstances [8]. An essential characteristic of decision tree routes is that they are exhaustive and mutually exclusive. The principle is as follows: a new set of training samples is generated from the original training samples, which generates a random composed of multiple decision trees, and the regression number uses the mean value to predict the results [5]. Every occurrence is replaced by a rule or a path.

3.2.1. Data pre-processing and evaluation volume. The whole data are separated into train data and test data, which contain 0.8 and 0.2 in random order respectively. Use Python to fit these data into the model to choose the number of decision trees and calculate the mean to make the final prediction. Here r^2, MAE, MSE, and RMSE are still applied to evaluate the model.

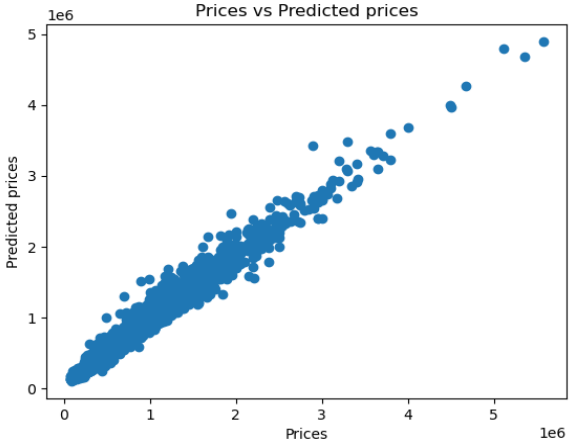

Figure 4. Linear relationship between predicted house prices and real house prices

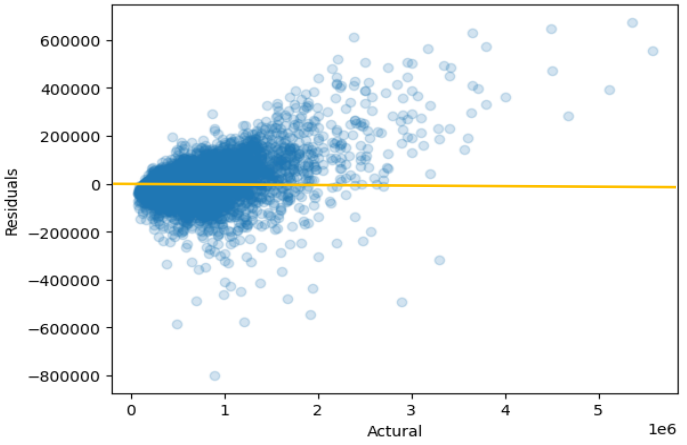

Figure 5. Relationship between true value and residual

3.2.2. Result. The first scatter plot shows the linear relationship between the actual price and the predicted price. It shows a good linear relationship in Figure 4. Still, r^2 presents a high level of 0.9759, so the degree of fit is good. Its MAE,MSE,RMSE are 30692.0389,3027220592.1187,55020.1835 respectively. The second plot indicates the relationship between the actual price and residuals. Although the actual price becomes higher than 2.5, the residuals in figure 6 show quite huge ups and downs, the density of these dots is very low, which means the quantity of them is small because the color is very light. The points with high density all fall around the residual of 0, so the model has better prediction

DOI: 10.54254/2753-8818/26/20241019

ability. This also shows that the Random Forest model is pretty accurate in predicting smaller values. Random Forest Regression can simultaneously process and classify numerical features and also reduce the risk of overfitting by averaging decision trees. However, if there is noise in the training data for classification or regression problems, the data set in the random forests is prone to overfitting.

3.3. XGboost Regression

The basic idea of XGboost is to train multiple weak decision trees in series to form a stronger decision tree. Each decision tree may not have a good classification effect, but multiple decision trees will give more accurate results [9]. The XGBoost algorithm combines several CART trees in the gradient descent direction in accordance with the loss function to minimize the loss function. It can also automatically utilize the CPU's multithreading for distributed learning and multi-core calculation, which can increase computational efficiency while ensuring classification accuracy and is appropriate for processing large amounts of data [10].

3.3.1. Process. The whole data are separated into train data and test data, which contain 0.8 and 0.2 in random order respectively. Use Python to fit these data into the model to add the trees, and the feature splits are constantly performed to grow a tree. Adding a tree each time is like learning a new function. Then it will fit the residual of the last prediction. Here r^2, MAE, MSE, and RMSE are also applied to evaluate the model.

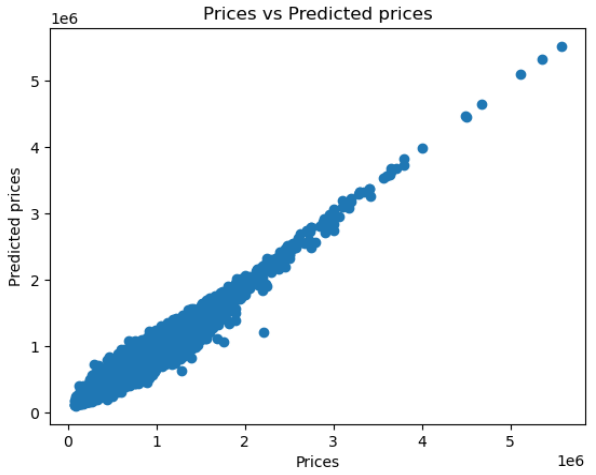

Figure 6. Linear relationship between predicted house prices and real house prices

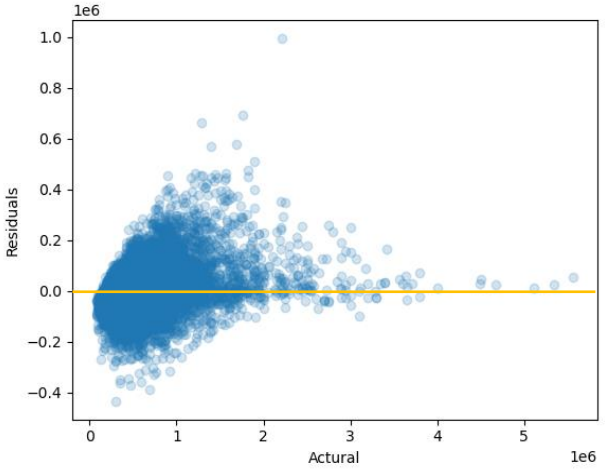

Figure 7. Relationship between true value and residual

DOI: 10.54254/2753-8818/26/20241019

3.3.2. Result. Figure 6 shows the linear relationship between the actual price and the predicted price. It shows a quite good linear relationship. Still, r^2 presents a quite high level:0.9535, so the degree of fit is good.The MAE,MSE,RMSE are 53433.8610,5833550734.8429 and 76377.6848 respectively. The second plot indicates the relationship between the actual price and residuals. When the actual price is in the range of 0 to 2.5, there is a large quantity of the does, because the color is very dark. Furthermore, as the price grows higher, the residuals are getting closer to 0. In addition, according to figure 7, points with high density all fall around the residual of 0, so the model has better prediction ability. This shows that the XGboost model is accurate in predicting values. XGboost Regression can efficiently handle large data sets with fast training speeds and predictive speed. It also has high accuracy and generalization capabilities to handle abnormal data and complex linear problems, while also reducing overfitting.

4. Conclusion

Each of the three models has its own characteristics. The principle of the Linear Regression model is the most simple and intuitive one, but the linear relationship is poorly fitted, the residual error is large, and the test result of the test set is not ideal. Random Forest Regression can process data quickly, the process is simple, and the trained data set has a higher degree of fit,which is 0.8172.However, the residual increases with the increasing number of data, and has a greater impact because of the existence of abnormal data. And the test set results are poor,which r^2 is just 0.6541.The XGboost Regression model is efficient and uses regular distributions to avoid large numbers of abnormal data. The linear relationship of its training set is clear and obvious. The residual distribution is relatively uniform and close to 0. The degree of fitting also becomes higher with the increase of the value. Test results in the highest score:0.8293. Therefore, the author believes that the XGboost Regression model is the best and most accurate one among the three machine learning models in predicting USA house prices.

This paper can be used as a reference for American students, investors, and other groups to buy a house and forecast the house price. Of course, there are some shortcomings in this study, such as the small independent variables and the lack of segmentation of urban geography. In the future, the detailed variables will be processed and the corresponding model with better predictive effect will be confirmed.

References

[1] Salim, L. Stelios, B. Christos, A. 2023, A comparative assessment of machine learning methods for predicting housing prices using Bayesian optimization, (Decision analytics journal,vol. 6), 100166.

[2] Lawrence, R.M. Yanru. 2009, Rent a house to buy a house in the United States. (Journal of resource and environment, vol. 23), pp. 3.

[3] Li, S. D. 2021, Prices based on multivariate linear regression forecast model. (Science and technology innovation, vol. 006), pp. 91-92.

[4] Hu, Y. X. Huang, Y. Wang, T. et al. 2018, Housing price trend analysis based on neural network prediction model: A case study of Haikou and Sanya. (Fujian computer, vol. 34), no. 12, pp. 2.

[5] Chen, S. P. Jin, S. P. 2016, Housing forecast based on random forest model. (Science and technology innovation and applications, vol. 4), pp. 1.

[6] Zhong, L. Y. Gao, S. L. 2017, Application of multiple linear regression model in housing price trend analysis and forecast. (Science and Technology Entrepreneurship Monthly, vol. 9).

[7] Zhang Re day. Housing forecast model based on multiple linear regression to improve . Journal of electronics, 2019 (4) : 3.

[8] Li, Y. Q. 2018, Housing forecast model based on random forest. (Journal of communication world, vol. 9), pp. 3.

[9] Zhang, J. Q. Du, J. 2020, Housing Price Prediction Model based on XGBoost and Multiple machine Learning methods. (Modern information technology, vol. 4), no. 10, pp. 4.

[10] Yang, G. J. Xu, X. Zhao, F. Q.2019, User rating prediction model based on XGBoost algorithm and its application. (Modern Library and Information Technology, vol. 3), no. 1, pp. 118-126.