Analysis of USA National Home Prices Based on Different Machine Learning Models

Yujie Li*

Sun Yat-sen University Cancer Center, Sun Yat-sen University, Guangzhou 510060, China

*liyj66@mail3.sysu.edu.cn

Abstract. Numerous nations rely heavily on the real estate industry, and changes in home prices have a big impact on people's quality of life. On this basis, house price prediction plays an important role in the economic field, e.g., making economic policy. Affected by thousands of potential factors, it is complicated to estimate the house price accurately. This study uses several machine learning models to build the relationship between 5 different factors of macro perspective with house prices in the US and managed to predict the real estate price. Among these models, the Decision tree model, KNN model, and Neural network model all perform high fitting effects and stable generalization activity. The SVR model is also suitable for this case. The article also indicates that the MLR model shows the worst fitting effect because of being limited in capturing the non-linear characters in datasets. Overall, these results provide accurate house price prediction models, which may be very valuable in real property sectors.

Keywords: Economy, house price prediction, machine learning.

1 INTRODUCTION

The global economy of many nations is heavily reliant on the real estate sector. The real estate sector is not only just a collection of properties, but also a dynamic sector including construction, buying, selling, and rental activities and each transaction in this sector [1]. It can affect the whole economy, catalyzing growth, and prosperity of economy by a series of activities, like the taxes paid on a transaction, the jobs created during construction or the influx of capital from real estate investments [2]. The majority of households in the US often have more worth in their homes than in their financial possessions [3]. Real estate has been and still is the foundation of wealth for millions of Americans and an essential means of transferring goods, services, and capital [4]. Thus, the focus on house prices is crucial in economy field. An increase in property values is a sign of increased wealth, which influences people's purchasing decisions [1]. Consumers tend to engage in riskier economic bets because they may feel more confident when their wealth is increasing. In general, they spend more for consumption and investment during the period of home values increasing as they are aware that they have assets to afford possible troubles [5]. Because consumer spending is a major component of the US economy, this spending boom could have a significant beneficial impact on

R. Magdalena et al. (eds.), *Proceedings of the 2024 9th International Conference on Social Sciences and Economic Development (ICSSED 2024)*, Advances in Economics, Business and Management Research 289,
https://doi.org/10.2991/978-94-6463-459-4_13

it. Consumer spending is expected to slow down if housing prices start to decline, which might be detrimental to the economy.

According to previous research, there are many factors that can influence home prices. These factors can be divided into three aspects. The first aspect includes micro factors, e.g., property location, upgrades and renovations, inspection report [6]. Secondly, macro factors, include a wide range of economic indicators, including the consumer price index (CPI), interest rates (Fed interest rate), population growth rate, GDP, unemployment rate, and income growth rate [7]. Thirdly, other factors include some policies, and mortgage interest tax [8]. Being influenced by so many factors, it is significantly complicated to predict the accurate house price in the real economy field.

Many academics established machine learning models to predict the price of a house based on the possible factors that could affect the price. Some scholars choose macro factors, for example, data from politics, society and economy fields to build the model for house price prediction. Others choose the inherent property of buildings as factors to predict the price, for example, the location, the number of bedrooms and so on. This research develops many different machine learning models to predict, including Multiple Linear Regression (MLR) model, Ensemble-learning method, Support Vector Regression (SVR) model, Decision-making tree model, and Neural Network method. In 2006, Fan et al. introduced and utilized the Decision Tree approach to investigate house prices in Singapore, illustrating the connection between housing property and house prices as well as important variables that could have an impact on the price of a home [9]. In 2009, Li et al. used the SVR approach to forecast Chinese real estate prices by examining a dataset containing five variables between 1998 and 2008 and indicated that another effective method for estimating real estate prices is the SVR model [10]. In 2014, Khamis et al. estimated house prices in New York by using the Multiple Linear Regression (MLR) model and the Neural Network model and compared the results. The estimation took into account various factors such as living area, age of the house, lot size, and number of bedrooms and bathrooms and this model was trained using 1047 houses as samples. The outcome demonstrates that the Neural Network model performs better than the MLR model [11]. In 2018, Yang et al. created an Ensemble-learning based house price prediction model and evaluated its performance against the GBDT, XGB, Random Forest, and Extra Trees algorithms. They discovered that Ensemble-learning can lead to better performance as well as increased prediction accuracy and stability [12]. There are also many groups that have predicted the house price by neural network algorithm, and all get high-level performance [13-15]. All the previous models have their own limitations and deficiencies, but they provide a variety of examples for house price prediction and model building, which have a positive significance in economy field.

This study uses different machine learning algorithms for models of house price prediction. The datasets from 1987 to 2023 of house prices and factors that may affect house prices in the US (like GDP, CPI and so on) are used to train these models. Then, the performances of these models are compared according to indicators RMSE, R^2 and Cross-validation. Finally, the results of these models are explained.

2 DATA AND METHOD

2.1 Data

The datasets used in this article are accessible from the Kaggle database, which include many indicators that may be related to house price in the US and the house price index of the US. Following are these indicators. US Mortgage Rates reflect the average interest rates on mortgage loans in the United States, mainly indicating the cost of borrowing for housing. The data contains Mortgage Rates in the US from 02.04.1971 to 26.10.2023 and the Mortgage Rates were recorded every week. Gross Domestic Product (GDP) represents the increase in monetary value that created by producing a wide range of goods and services produced in the US over a given time frame. At the same time, it also indicates the income derived from the production and the sum that is spent on these goods and services. This indicator is one of the most important economic indicators and it can demonstrate the economic health and activity of one country. The data contains the Gross Domestic Product (GDP) in the US from 01.01.1947 to 01.07.2023 and was recorded every three months. Unemployment Rates correspond to the percentage of the labor force that was unemployed at that time, which is an important sign of the health of the labor market and economic stability. The data contains unemployment rates in the US from 01.01.1948 to 01.09.2023 and was recorded every month. In the US, the Federal Open Market Committee (FOMC), which determines the target interest rate range, sets the FED Funds Rate. Using this rate, commercial banks lend and borrow excess reserves from one another overnight. This rate is a primary tool for monetary policy, influencing borrowing costs and, subsequently, overall economic activity. The data contains the FED Funds Rate in the US from 01.07.1954 to 01.09.2023 and was recorded every month. Population Growth Rate indicates the annual population growth rate and reflects the change in population caused by births, deaths, and migration, which offers insights into demographic trends and has implications for the labor force, consumer markets, and social services planning. The data contains the Population Growth rate in the US from 01.01.1961 to 01.01.2022 and was recorded every year. Consumer Price Index (CPI) shows how prices for goods and services that consumers pay have changed on average over time, which is a key indicator for assessing inflation or deflation. The CPI indicator can influence consumer's behavior on spending and economic policy decisions. The data contains Consumer Price Index (CPI) in the US from 01.01.1947 to 01.09.2023 and was recorded every month. Case-Shiller Housing Price Index (USA) measures changes in the prices of residential real estate properties over time and it is crucial for assessing the state of the housing market, providing information about the state and patterns of the US housing market. The data contains Case-Shiller Housing Price Index in the US from 01.01.1987 to 01.07.2023 and was recorded every month.

Here, there are 6 features provided to predict the house price (Case-Shiller Housing Price Index) in the US, which mainly focus on the macro aspect. Due to the time points that were recorded in these features being different, one started to carry out data cleaning and temporal alignment at beginning. The data that was recorded earlier than 01.01.1987 and later than 01.01.2021 was deleted. After observing, the scales of

different features vary significantly. Hence, Standardization was performed to bring different features to a common scale, ensuring that they have comparable magnitudes. Afterward, one analyzed the correlation of these features with house prices. The results show that the correlation between GDP and house price and the correlation between CPI and house price are positive and high, with 0.96 and 0.94 separately. The correlation between FED Funds Rate and house price, between US Mortgage Rates and house price and between population-growth-rate and house price are negative and medium, with -0.68, -0.81 and -0.79 respectively. The correlation between Unemployment Rates and house price are low, only -0.22. So, one deleted the feature of Unemployment Rates in the following model building.

2.2 Models

This study mainly used 5 different machine learning models to train this dataset and predict the house price, following are detailed information. Multiple linear regression (MLR) model, one of the models that establishes a linear relationship between several independent variables and a dependent variable is known as multiple linear regression. To achieve the best fitting effect, it is crucial to minimize the error between the predicted and actual values during the fitting process. It would be realized by adjusting the model coefficients. The basic equation used in the model building includes an intercept term and coefficients for each independent variable and least squares methods are used to find the optimal combination of coefficients. Then new datasets are applied to test the accuracy and effect of fitting in this linear regression model.

The K-Nearest Neighbors (KNN) algorithm is the most basic and simple method to deal with classification and regression problems. The principle of this method is to predict the values of new data by known points in train samples which are the nearest in distance defined by users. In this paper, some parameters are searched to get the best fitting effect. One of the most important parameters is N_neighbors. It indicates the number of data points which are counted for calculating new data values. Here the values of N_neighbors from 1 to 49 are searched and the best N_neighbors is 4. Another parameter in KNN algorithm is weights, which can specify the weight of neighbors. Uniform weight ('uniform') means that each nearest neighbor contributes equally to the prediction. Distance Weight ('distance') means that closer nearest neighbors have a greater influence on the prediction. In this prediction, distance Weight can get a better effect. 'P' is also a parameter that can affect the prediction, which means how to calculate the distance. The Manhattan distance is represented by the value of "p" = 1 and the Euclidean distance is represented by the value of "p" = 2. Here one uses Manhattan distance to do prediction.

The principle of the decision tree is to separate sample datasets into several different regions by a series of leaf nodes in the tree. At each leaf node, it is necessary to select features and determine thresholds to realize classification and calculate the average variance of the target variable. The model is successful when the average variance of the target variable in every region is the least. When predicting new samples, it can separate these new samples into different parts and realize regression model building. In this article, two parameters are searched to obtain the best fitting effect. Firstly, maximum

depth (‘max_depth’) is searched, which means the maximum depth or length of the tree. It sets a limit on the number of nodes that extend from the tree's root to its furthest leaf node. In this prediction, ‘max_depth’ equals to 9. The tree must have a suitable value for "max depth" in order to avoid overfitting the training set and becoming overly complex. Minimum Samples per Leaf (min_samples_leaf), or the least amount of samples needed to form a leaf node, is the second parameter. It is also important for optimizing the performance of the decision tree model in regression tasks. In this fitting, one chooses 'min_samples_leaf' equal to 3.

The principle of SVR is to find a hyperplane with an epsilon-insensitive tube to fit the sample dataset, aiming to minimize the prediction error by defining that errors within this tube are considered acceptable. In this model, nearly all crucial data points are used to define the tube's position and width. Afterward, a kernel function, such as the Radial Basis Function (RBF) kernel, is employed to transform input features into a higher-dimensional space to realize fitting. In this article, the Radial Basis Function (RBF) kernel was used to introduce a non-linear relationship to deal with data with complex patterns. Two parameters are searched to obtain the best fitting effect, that are epsilon and C values. Epsilon is an important parameter which means the width of the epsilon-insensitive tube and errors within this tube are considered acceptable. In this model, the epsilon equals to 0.01. The parameter ‘C’ represents the regularization parameter, which is a crucial hyperparameter that realizes the balance between the complexity of the model and fitting the training data accurately. In this model, the C values are equal to 4.

Neural networks are machine learning models composed of neurons and a hierarchical structure including input, hidden, and output layers. In this model, sample datasets are trained through forward and backward propagation processes. Neurons receive inputs, generate outputs through forward propagation processes and optimise algorithms by gradient descent to adjust weights and biases to minimize the loss function. Activation functions in hidden layers are used to introduce non-linearity to get better fitting effect. In this model, hidden_layer_sizes, the choice of activation functions, and learning rate are parameters that need to be defined. Here, one use hidden_layer_sizes= (100,50,100), activation='relu', learning_rate_init=0.01.

After model building, it is also important to evaluate the fitting effect by some indicators. Here are some Model Evaluation Metrics used in this article. Root Mean Square Error (RMSE): A popular metric for examining a model's impact is root mean square error (RMSE). In regression analysis, it shows how accurate a predictive model is. A lower RMSE denotes a better performance. It is the square root of the average values of the squared differences between predicted values and actual values. Here is the formula of RMSE:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{n}} \tag{1}$$

where, n is the quantity of observations, y_i is the dependent variable's actual value for the observation, $\widehat{y_i}$ is the dependent variable's predicted value for observation. R^2 is a metric used to evaluate the fitting effect of a regression model, which is also known as the coefficient of determination. It indicates the proportion of the variance between

the predicted models and actual values. The values of R-squared (R^2) ranges from 0 to 1 and a higher R^2 indicates a better fitting. Here is the formula:

$$R^2 = 1 - \frac{\sum i\,(\widehat{y_i} - y_i)^2}{\sum i\,(\overline{y_i} - y_i)^2} \tag{2}$$

Here, n is the quantity of observations, y_i is the dependent variable's actual value for the observation, $\widehat{y_i}$ is the dependent variable's predicted value for observation, $\overline{y_i}$ is the mean of the actual values of the dependent variable.

The numerator indicates errors caused by prediction in model. The denominator indicates a standard if the predicted values equal to the average of actual values. One statistical method for evaluating a fitting model's performance and preventing overfitting is cross-validation. This model divides the training dataset into five equal portions and uses k-fold cross-validation. Next, it achieves training on k-1 folds and tests on the remaining fold in each iteration of this Cross-Validation procedure. Then every R-squared (R^2) metric is averaged over the k iterations to obtain a more reliable estimate and one can finally assess the overall performance and reliability of a predictive model.

Table 1. The fitting effects of 5 machine learning models evaluated by RMSE, R^2 and R^2 in Cross-validation.

	MLR	KNN	Decision-tree	SVR	Neural Network
RMSE	0.2772	0.0568	0.0438	0.0760	0.0499
R^2	0.9284	0.9970	0.9985	0.9952	0.9977
R^2 in CV	0.9180	0.9974	0.9961	0.9917	0.9970

3 RESULTS AND DISCUSSION

The datasets are divided into train dataset and test dataset, including 5 features (US Mortgage Rates, GDP, FED Funds Rate, Population Growth, CPI) and house price index (Case-Shiller Housing Price Index). One chooses 80% of sample datasets to train the model and 20% of the datasets to test the fitting effect using 5 different machine learning methods. To test the efficiency and accuracy of these models, one used RMSE, R^2, and R^2 in Cross-Validation to evaluate these models. The results are summarized in Table 1 and illustrated in Fig. 1 and Fig. 2.

The RMSE value of Multiple Linear Regression (MLR) model is relatively high, with 0.2772, which means that the fitting effect of Multiple Linear Regression (MLR) model is unsatisfactory and the error in MLR predictions is high. Being consistent with RMSE, R^2 of this prediction is also relatively low, only equals to 0.9284, which means the variance of explaining the target variates is high. The R^2 result of cross-validation is also low, only equals to 0.918, which means that the fitting effect in training dataset is unstable. The K-Nearest Neighbors (KNN) algorithm is one of the simplest algorithms in machine learning. However, the fitting effect of this model is high, which is verified by low RMSE value of KNN prediction (only 0.0568). At the same time, R^2 value of KNN prediction is 0.997, which is remarkably high, indicating that KNN can

explain a significant portion of the target variable's variance. The R^2 value in cross-validation is 0.9974, which is also very high. That means that this prediction has strong generalization ability. Decision Tree prediction results also show well-fitting performance. RMSE in decision tree fitting is 0.0438, which is very low, indicating that the Decision tree has a very good fitting effect. R^2 value of this model is 0.9985, which is very high, suggesting that the variance in this prediction is low and stable. R^2 in cross-validation (0.9961) is also high, which means that the fitting effect of decision tree has a strong generalization ability.

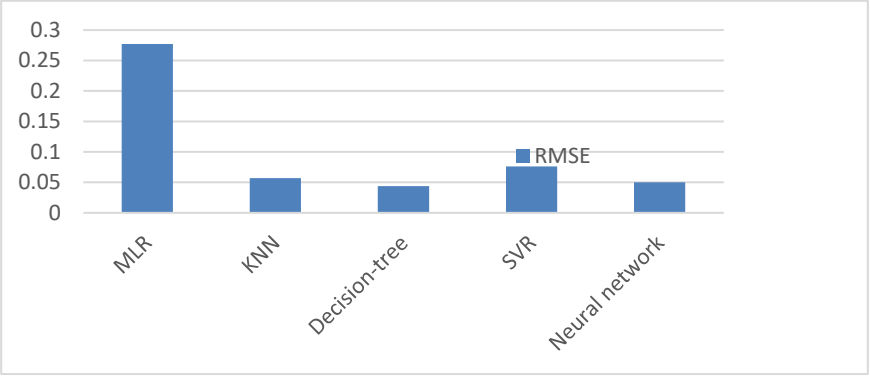

Fig. 1. The histogram graph for RMSE values comparison of 5 models (Photo/Picture credit: Original).

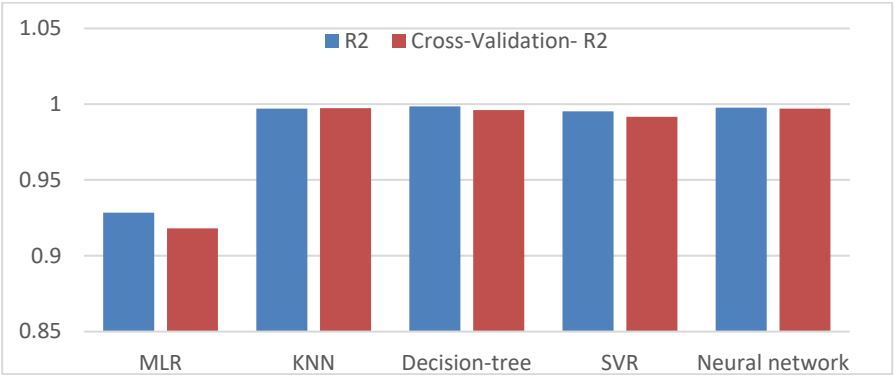

Fig. 2. The histogram for R2 and R2 in Cross-validation values comparison of 5 models (Photo/Picture credit: Original).

The results of Neural Network are the same as Decision Tree, with low RMSE (0.0499), very high R^2 (0.9977) and high R^2 in cross-validation (0.997). All means that this Neural Network model prediction can show very good performance and strong generalization ability. For support vector regression (SVR) model, RMSE value is 0.076 and R^2 value is 0.9952, which means that this method also shows relatively good fitting effect. R^2 value in cross-validation is 0.9917, indicating that this model also has good generalization performance.

Overall, by comparing the results of the 5 models, one can find that the Decision tree model prediction can get the best fitting effect, at the same time, Neural Network prediction and KNN prediction can also get very good performance being similar to Decision tree model. The following model which can also be chosen is SVR model, although the fitting effect is not as good as Neural Network, KNN and Decision tree model. These models all realize very high fitting performance and are suitable on house price prediction in this case. On the contrary, among these models, MLR prediction shows the worst fitting effect. After analysis, one concludes the reasons why MLR shows the worst fitting effect. It is evidence that the relationship between target variables and house prices shows a non-linear fitting property. In other models, such as KNN, SVR, Neural Network, and Decision tree, they are easy to capture the non-linear characters in datasets, especially in Decision tree models with non-linear fitting capabilities, allowing them to adapt better to the data. Whereas, in MLR model, it is limited to capturing non-linear relationships in the data.

4 LIMITATIONS AND PROSPECTS

This study attempts to forecast US home prices using various machine learning algorithms. Here, one mainly utilizes 5 different features in macro aspect, including GDP, population growth rate, CPI, FED Funds Rate, US mortgage rate as variates to construct models using various techniques in order to forecast the price of homes. Besides, one gets a very good fitting effect in this article. In reality, a variety of factors, such as location, culture, and policies, affect how much a house costs. Therefore, predicting the real estate market's price is difficult. If one can obtain an increasing number of datasets from various perspectives, the prediction model's accuracy might improve. In addition, the datasets one used here have only 421 time points, which is limited. The limited time points of this data may also affect the performance of house price prediction. Meanwhile, the research district is only restricted to the US, and one only carries out the macro-level analysis, which may affect the universality of these models. In the future, it is necessary to introduce more datasets and features from different aspects to improve the prediction, because more data points can also improve the accuracy of the prediction model. At the same time, datasets from different countries or districts can also be applied to train the model, which is necessary to test the university of house price prediction model. Finally, the more complex model may also need to be applied to deal with complex scenery, for example, random forest methods.

5 CONCLUSION

To sum up, this study combines five different factors that may influence the price of a house to successfully build several machine learning models to predict the price of a house in the United States. Most of these models show very good fitting performance, including Decision tree model, KNN model, Neural Network model. At the same time, there are also some limitations for the prediction, including limited dataset numbers and time points and restricted areas. In the future, more datasets including other factors may

help to generate more accurate prediction models. Overall, this paper provides accurate house price prediction models, which may be very valuable in real property sectors and the economy field.

REFERENCES

1. Zhao, C., Liu, F.: Impact of housing policies on the real estate market-Systematic literature review. Heliyon (2023).
2. Mallick, H., Mahalik, M. K.: Constructing the economy: the role of construction sector in India's growth. The Journal of Real Estate Finance and Economics, 40, 368-384 (2010).
3. Grinstein-Weiss, M., Key, C., Carrillo, S.: Homeownership, the great recession, and wealth: Evidence from the survey of consumer finances. Housing Policy Debate 25(3), 419-445 (2015).
4. Gotham, K. F.: The secondary circuit of capital reconsidered: Globalization and the US real estate sector. American journal of sociology, 112(1), 231-275 (2006).
5. Fan, Y., Yavas, A.: How does mortgage debt affect household consumption? Micro evidence from China. Real Estate Economics, 48(1), 43-88 (2020).
6. Zietz, J., Zietz, E. N., Sirmans, G. S.: Determinants of house prices: a quantile regression approach. The Journal of Real Estate Finance and Economics, 37, 317-333 (2008).
7. Baharuddin, N. S, Isa, I. N. M, Zahari, A. S. M.: Housing price in Malaysia: the impact of macroeconomic indicators. Journal of Global Business and Social Entrepreneurship (GBSE), 5(16), 69-80 (2019).
8. Sommer, K., Sullivan, P.: Implications of US tax policy for house prices, rents, and home-ownership. American Economic Review, 108(2), 241-274 (2018).
9. Fan, G. Z., Ong, S. E., Koh, H. C.: Determinants of house price: A decision tree approach. Urban Studies, 43(12), 2301-2315 (2006).
10. Li, D. Y., Xu, W., Zhao, H., Chen, R. Q.: A SVR based forecasting approach for real estate price prediction. 2009 International conference on machine learning and cybernetics, 970-974 (2009).
11. Model, A. B. K., Kamarudin, N. K. K. B.: Comparative Study On Estimate House Price Using Statistical And Neural Network. International journal of scientific and technology, research, 3(12), 126-131 (2014).
12. Cao, B., Yang, B.: Research on ensemble learning-based housing price prediction model. Big Geospatial Data and Data Science, 1(1), 1-8 (2018).
13. Kostic, Z., Jevremovic, A.; What image features boost housing market predictions? IEEE Transactions on Multimedia, 22(7), 1904-1916 (2020).
14. Wang, J., Hu, S., Zhan, X., et al.: Predicting house price with a memristor-based artificial neural network. IEEE Access, 6, 16523-16528 (2018).
15. Ahmed, E., Moustafa, M.: House price estimation from visual and textual features. arXiv preprint arXiv:160908399 (2016).