# Research on the Prediction of US House Prices Based on Machine Learning

Yunling He*

Leeds joint school, Southwest Jiaotong University, Chengdu, 611756, China

*Corresponding author: sc20yh@leeds.ac.uk

**Abstract.** With the increase in the standard, houses have become an indispensable part of people's lives. However, purchasing houses needs to consider the prices. Therefore, this paper uses machine learning to predict US house prices. Based on the random forest and linear regression method used in this research. It is found that the use of the former has a good effect on predicting multivariate nonlinear relationships in house prices. After comparing the results of actual prices and predicting prices, residents can select different room structures by contrasting them with experimental types as the numerical results are basically accurate. This research provides theoretical value to the literature on housing price prediction and brings different enlightenment to policymakers, regulators, and investors. As a result, this research plays a key role in maintaining the order of the real estate market and helping people make good choices.

**Keywords:** House Prices, Prediction, Random forests, Linear regression

## 1. Introduction

According to Zillow Research [1], the national unemployment rate fell below the natural unemployment rate, just 4.3 percent in 2017. And the 30-year fixed-rate mortgage's typical interest rate (the most common loan product type for U.S. home buyers) is 4%. A very strong job market, coupled with historically high housing affordability, stimulates the demand for people to buy a house. For most people, there are two key factors to stimulate them to get the need to purchase houses. One is landing a steady job that makes people can afford the monthly mortgage payments on a home. Another is low mortgage interest rates also make benefit to relieve the pressure of buying a house by installments.

There are many theoretical papers that have already conducted some research on the house price prediction. Christopher and Todd (2010) used two periods of economic prosperity to examine the relative role of economic fundamentals and market psychology in explaining house price dynamics in the United States [2]. Changchun and Hui (2018) compared the linear regression model and random forest model to predict house prices [3]. Quang et al. (2019) found that house prices strongly correlate to location, area, and population and used them in the extreme gradient boosting model, hybrid regression model, stacked generalization model, etc. to estimate the house price [4].

In addition, some research introduces some machine learning methods in different areas. Xiaolan et al (2022) used machine learning to determine the energetic material's qualities in advance [5]. Procedia Computer Science (2021) used machine learning to predict the trend of the stock market [6]. Susmita et al (2022) found that machine learning can better fit when designing the structures of rubble-mound breakwaters [7].

Thus, this research aims to adopt the linear regression model and random forest model to predict US house prices in size and layout. This study has played a great role in both theory and practice. It enriches the literature on machine learning prediction of housing prices and gets revelation to policymakers, regulators, and investors.

The rest of the paper is organized as follows: Section 2 introduces data and methods. Section 3 shows the results. Section 4 concludes.

# 2. Data and Method

## 2.1 Data and Processing

A dataset called "Housing Price in the USA" has more than 80000 data points and 18 variables that affect house prices from different aspects are as geographical location and internal structure. Table 1 shows the details of each attribute. These data were used to predict the price of different single houses in the USA.

The next procedure was to handle outliers. Many anomalies need to be dealt with. The following are some of the characteristic engineering procedures performed to clean up the dataset:

Firstly, replace the value of bedrooms, bathrooms, and price with integers. Secondly, change the attribute “statezip” from object to int. Thirdly, find the error values in the dataset. It is easy to find the errors that houses where prices are 0, no bedrooms or bathrooms in houses, the most expensive or cheapest prices, and the widest square footage of the home or lot. Fourthly, drop error datapoint.

**Table 1.** List of Attributes

| Attribute Name | Data Type | Description |
|---|---|---|
| date | object | Record time |
| price | float64 | House price |
| bedrooms | float64 | Number of bedrooms |
| bathrooms | float64 | Number of bathrooms |
| sqft_living | int64 | Square footage of the home |
| sqft_lot | int64 | Square footage of the lot |
| floors | float64 | Total floors in the house |
| waterfront | int64 | House which has a view of a waterfront |
| view | int64 | Has been viewed |
| condition | int64 | How good the condition is Overall |
| sqft_above | int64 | Square footage of house apart from the basement |
| sqft_basement | int64 | Square footage of the basement |
| yr_built | int64 | Built year |
| yr_renovated | int64 | The year when the house was renovated |
| street | object | The street where the house is located |
| city | object | City where the house in |
| statezip | object | Zip code |
| country | object | USA |

## 2.2 Method

### 2.2.1 Machine Learning

The objective of machine learning, a branch of artificial intelligence and computer science, is to replicate human learning processes using data and algorithms. In 1997, a man named Mitchell put forth a clear description [8]. If a computer program improves its performance on a task in T by using experience E (Experience), then talk about T and P, and the program learns E. Assume that P (Performance) is used to evaluate the performance of a computer program on a specific type of task T (Task).

### 2.2.2 Package Selection

This paper’s main packages imported include Pandas, NumPy, Scikit-learn, and Matplotlib. Here are four main packages in python:

#### 2.2.2.1 Pandas

The name Pandas comes from a combination of the terms Panel Data and data analysis. In this research, it plays an important role in the following aspects:

1. Provide a simple and efficient DataFrame object with a default or custom labels.
2. Load the dataset through a CSV file and convert it into the processable objects.
3. Data normalization operations and missing value handling can be easily implemented.
4. Convenient to add, modify, or delete data columns in DataFrame.
5. Provides a variety of ways to work with datasets, such as building subsets, slicing, filtering, grouping, and reordering.
6. Integration with other libraries is implemented like Scipy, Scikit-learn, and Matplotlib.

#### 2.2.2.2 Numpy

The name NumPy stands for Numerical Python, an array object library with several dimensions (also known as Nd arrays) and a selection of functions for handling these arrays. Users may operate mathematically and logically related operations on arrays using the NumPy library. In this study, NumPy uses two models to measure the mean root square errors.

#### 2.2.2.3 Scikit-learn

Scikit-learn plays a key role in machine learning. It is a machine learning library in Python, based on NumPy, SciPy, Matplotlib, and other data science packages, covering almost all aspects of sample data, data preprocessing, model verification, feature selection, classification, regression, and so on in machine learning, and the functions are highly potent. In this project, it gives various assistance in the following ways. Firstly, *LabelEncoder ()* is applied to train and encode the dataset. Secondly, *train_test_split* divides the dataset into the training set and the testing set in a certain percentage. Thirdly, use *LinearRegression ()* and *RandomForestRegressor ()* to set the models predicting house prices. Fourthly, calculate the mean absolute error and root mean square error by employing *mean_absolute_error ()* and *mean_squared_error ()* functions.

#### 2.2.2.4 Matplotlib

Matplotlib is one of the most popular data visualization software packages in Python, supports cross-platform operation, is a commonly used 2D drawing library in Python, and provides some 3D drawing interfaces. Matplotlib is often used with NumPy and Pandas and is one of the indispensable and important tools in data analysis. In this research, matplotlib is used to draw the figures that show the comparison of actual prices and predicted prices.

### 2.2.3 Package Selection

Performing data analysis is a critical step in creating a regression model. By this means, it can help choose appropriate machine learning approaches by finding the connections between different variables.

For the models to understand the patterns more quickly, the number should be treated appropriately before building them. In general, numerical values are standardized while categorical data are labeled encoded. Then, split data into two groups individually variable input and target class. The group 'variable input' elements delete the columns containing 'date', 'street', 'country', 'price', 'statezip', and 'yr_renovated'. The group 'target class' only comprises 'price', the object we research. Finally, the dataset is divided into a training set and a test set with a ratio of 10:3 by utilizing the scikit-learn package.

The evaluation is performed by "score" in scikit-learn. The return of this function is the coefficient of determination [9]. The coefficient of determination is denoted as $R^2$. The amount of variation in induced variables can be stated by the dependence on arguments. If the value of $R^2$ is bigger, the fit is better. The upline of the coefficient of determination is 1 which means if the value gets closer to 1, the model can better explain the variations of the outputs with different input variables [9].

Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are also taken utilities to estimate the degree of the two models. The mean of the absolute errors between the actual and expected values is known as the MAE. RMSE, which reflects the difference's sample standard deviation between the predicted value and the observed value that demonstrates the dispersion of the sample when doing the nonlinear fitting. The lower the RMSE, the better. But compared to RAE, RMSE is greatly affected by the outliers. It may influence the judgments of the two models.

#### 2.2.3.1 Linear Regression

A straight line is created by using the linear regression model to forecast or display a correlation of direct proportionality between the predictor parameters and the reliant variables. [10].

An outcome variable is explained by linear regression using a single predictor variable. Here is a straightforward linear regression equation: [4]:

$$y_i = \alpha + \beta x_i + \epsilon_i \quad (1)$$

Where x is the forecaster or independent variable used to predict Y, y is the relying factor, β is the regression coefficient, α is the intercept (expected average value of home values whenever a given independent variable is 0), and ϵ is the error term, which takes into account the randomness that the models are unable to explain.

#### 2.2.3.2 Random forest

A type of integrated model with numerous decision trees is called Random Forest. If the classification job receives more fresh input samples, each decision tree in that forest will permit to judge and classify separately. Each classification outcome from a decision tree is unique. The decision tree's categorization findings with the highest quantity will be used as the basis for the random forest's conclusion.

The following four phases can be used to illustrate the random forest method [11]:

Step 1: Take random samples and train decision trees. Construct a train set with N elements. Every time take one object out and put it back. Repeat this process n times and hence formed n samples. As the specimens at the decision tree's root nodes, the chosen N samples are utilized to train the decision tree.

Step 2: When the sample has M attributes, choose m features randomly in them without replacement. Use a characteristic to split the node that offers the best split in terms of the objective function, such as maximizing information gain.

Step 3: Duplicate step 2 until every node is split.

Step 4: Constitute a random forest. Aggregate the results decided by trees and make a conclusion by majority vote.

## 3. Results and Discussion

In this section, this paper conducts the analysis by comparing the results got by the linear regression and random forest.

### 3.1 Result on Linear Regression

This paper uses the linear model class provided by sci-kit-learn. The result of this model is shown in figure 1. It suggests the results are proper. The coefficient of determination training is about 0.588 and the coefficient of determination test is about 0.568. The former result is higher than the latter means that the predicting value meets the expectations. However, the coefficient of determination training still has a huge distance from 1. Therefore, it may not a fine model to predict the house price.

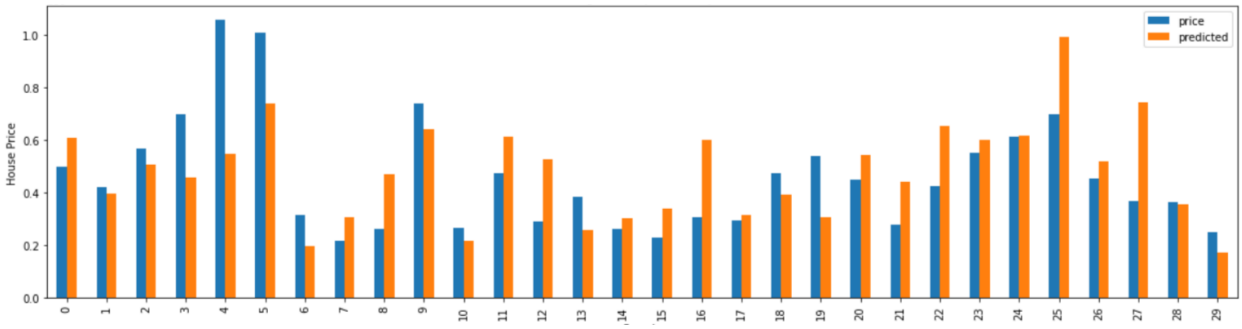


**Fig 1.** Some comparisons of price and predicted results in linear regression

### 3.2 Result on Random Forest

This paper used the Decision Tree Regressor class and Random Forest Regressor class by sci-kit-learn. As the amount of data is not huge, the max quantity of the decision tree is not set.

However, some parameters are added to initialize the Random Forest regarding random forest subsets: Set max_depth = 18, which restricts these trees can only go to 18.

As shown in figure 2, the model implemented with high precision and the coefficient of determination training is 0.946. This figure is very close to 1 which means the data has a high fitting degree. This value proves that Random Forest is an appropriate approach to predicting house prices.

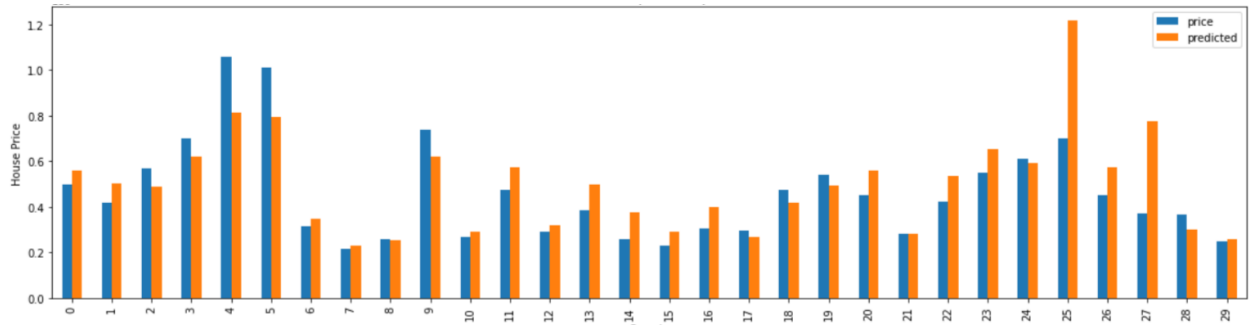


**Fig 2.** Some comparisons of price and predicted results in random forests.

### 3.3 Comparison on the efficiency of two models

The coefficient of the training set and the testing set in different models are presented in Table 2. MAE and RMSE have also evaluated that shown in Table 3. These results show that the two models used in this paper properly predict house prices. But by comparing the two models, it is obvious that random forest gets a better effect. For this reason, random forest is able to handle nonlinear information. In this prediction, multiple factors should pay attention to, but linear regression may get a nicer result if only one variable. As these statistics are influenced by a single value and the unit price of the house is a huge number, it is difficult to use it as a criterion for judging whether a model is good or bad. Different methods have different errors, if these obstacles can be solved in the future, the predicting prices can be more precise.

**Table 2.** Predicting Results of coefficient of determination

| Model | Coefficient of train | Coefficient of test |
|---|---|---|
| Linear Regression | 0.5885 | 0.5679 |
| Random Forest | 0.9461 | 0.6456 |

**Table 3.** Predicting results of MAE and RMSE.

| Model | Mean Absolute Error | Root Mean Square Error |
|---|---|---|
| Linear Regression | 155850.6793 | 225577.9324 |
| Random Forest | 47530.6423 | 204315.9175 |

## 4. Conclusion

This paper uses machine learning to predict US house prices and obtain three main conclusions. Firstly, the linear regression and the random forest models show striking contrasts. For this kind of research with a large number of variables, the predicted performance of random forest is greater than linear regression obviously. Secondly, most forecasts are higher than actual prices, implying that it is a suitable time to purchase houses at present. Thirdly, some predictions still get huge differences, which illustrates that the price-performance ratio is not high.

These findings are of great significance in two respects. One is enriching the literature on house price prediction theoretically, and the other is getting revelation to policymakers, regulators, and investors. For regulators, it facilitates the government's rational supervision of the housing market, resulting in more reasonable housing prices and more stable real estate market growth. According to adjustments in the market and policy environment, regular people can more intuitively understand changes in real estate.

## References

[1] Zillow Research. Why is there so much demand for home ownership? June 28, 2017. Retrieved on July 10, 2022. Retrieved from: https://www.zillow.com/research/why-is-home-buying-demand-high-15781/

[2] Christopher J. Mayer and Todd Sinai, U.S. House Price Dynamics and Behavioral Finance, Semantic Scholar, 2010, Pages 262-295, https://www.semanticscholar.org/paper/U-.-S-.-House-Price-Dynamics-and-Behavioral-Finance-Mayer/a07de769293d285524ec7a0b8c1370e4f3795086#related-papers

[3] Changchun Wang, Hui Wu, A new machine learning approach to house price estimation, BISKA, Volume 4, 2018, Pages 165-171. http://dx.doi.org/10.20852/ntmsci.2018.327

[4] Quang Truong, Minh Nguyen, etc., Predicting House Prices Through Improved Machine Learning Techniques, Procedia Computer Science, Volume 174, 2020, Pages 433-442, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.06.111.

[5] Tian Xl, Song Sw, etc., Machine learning guided property prediction of energetic materials: Recent advances, challenges, and perspectives, Energetic Materials Frontiers (2022), DOI: https://doi.org/10.1016/j.enmf.2022.07.005.

[6] Abdulhamit Subasi, Faria Amir, etc., Stock Market Prediction Using Machine Learning, Procedia Computer Science, Volume 194, 2021, Pages 173-179, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2021.10.071.

[7] Susmita Saha, Satyasaran Changdar, etc., Prediction of the stability number of conventional rubble-mound breakwaters using machine learning algorithms, Journal of Ocean Engineering and Science, 2022, ISSN 2468-0133, https://doi.org/10.1016/j.joes.2022.06.030.

[8] CMU. Machine Learning, Tom Mitchell, McGraw Hill, 1997. Retrieved on July 30, 2022. Retrieved from: https://www.cs.cmu.edu/~tom/mlbook.html#:~:text=Mitchell%2C%20McGraw%20Hill%2C%201997.%20Machine%20Learning%20is%20the,improve%20automatically%20through%20experience.%20Applications%20range%20from%20datamining

[9] CSDN. In the Scikit-learn library, the meaning of the score function used in regression performance evaluation is explained in detail. August 3, 2021. Retrieved on July 15, 2022. Retrieved from: https://blog.csdn.net/m0_48520385/article/details/119354081

[10] [Learn Data Sci. Predicting Housing Prices with Linear Regression using Python, pandas, and stats models. Retrieved on July 20, 2022. Retrieved from: https://www.learndatasci.com/tutorials/predicting-housing-prices-linear-regression-using-python-pandas-statsmodels/

[11] Easy AI. Random forest – Random Forest. March 8, 2019. Retrieved on July 24, 2022. Retrieved from: https://easyai.tech/en/ai-definition/random-forest/#google_vignette