Proceedings of the 2023 International Conference on Management Research and Economic Development
DOI: 10.54254/2754-1169/20/20230199

Research on the Influencing Factors and Prediction of Housing Prices Based on Regression Analysis --- Taking California as an Example

Kunlong Zhang[1,a,*]

[1]*University of Illinois at Urbana Champaign*
a kunlong2@illinois.edu
**corresponding author*

Abstract: A house is a necessity for everyone's life. But in today's world, with the continuous reduction of land area, the continuous increase of population, and the continuous maturity of the real estate industry, people have to consider more carefully what factors are the most important in the house. Based on the background of housing prices in California, this paper uses linear regression, random forest and principal component analysis to determine which variables have the greatest impact on housing prices. The reason for using three methods is to obtain more accurate results. According to the results, linear regression shows that income is the most relevant variable to housing prices. Random Forest shows an R square of 75.7%, meaning the predictions fit the data fairly well. Principal component analysis also shows that income is the most important variable. At the same time, house prices will be predicted based on the data obtained, which is about 116598.48588. Some suggestions for those who want to buy a house - which factors of the house are more important. Second, this article will also analyse how to face the rising housing prices.

Keywords: linear regression, correlation coefficient, random forest, principle component analysis, housing price prediction

1. Introduction

One of a person's most valuable possessions is their home. A house is the warmest and the foremost location where one feels quite secure. One can retain their possessions and other priceless items at home. If it is tastefully designed, it can be both the safest and best place to be. This aids in preventing theft of one's goods. Second, someone who is worn out from work or other activities will rush back home to unwind. One of the best places to get away from stress and worry is at home. The ideal place to spend your retirement years is at home. Thirdly, a home offers total freedom that might be appreciated. People are free to live the life they want and pursue their passions. No limitations will apply to it. To summarize, buying a home necessitates considerable thought. This paper's primary goal is to investigate, using a variety of techniques, which factors have a stronger impact on the price of homes. In order to identify the factors that have the most effects on housing prices, the data should first be analyzed using linear regression, correlation coefficient, random forest, and principal component analysis. then look at the reasoning behind it. Finally, some recommendations for addressing the problem of excessive housing prices are made.

DOI: 10.54254/2754-1169/20/20230199

2. Methodology

2.1. Source of Data

Information from the California census of 1990 is included in the data. The data is from Kaggle. Longitude, latitude, median income, housing median age, number of rooms, number of bedrooms, population, households, median housing price, and proximity to the ocean are the 10 variables that make up the data, with the median housing price serving as the dependent variable and the other 9 as independent variables.

2.2. Data Processing

Table 1: Overview of each variables.

	Longitude	Latitude	House median age	Total rooms	Total bedrooms	population	households	Median income	Median house value	Ocean proximity
0	-122.23	37.88	41	880	129	322	128	8.3253	452800	Near Bay
1	-122.22	37.88	21	7099	1108	2401	1138	8.3014	358500	Near Bay
2	-122.24	37.85	52	1467	190	498	177	7.2574	352100	Near Bay
3	-122.25	37.85	52	1274	235	558	219	5.6431	341300	Near Bay
4	-122.25	37.85	52	1627	280	565	259	8.8462	342200	Near Bay

According to the statistical data and Table 1, all the data are supported by enough samples, and there are no invalid attributes, so no cleaning is required.

2.3. Models

2.3.1. Linear Regression and Correlation Coefficient

A statistical analytic technique used to determine the interdependent quantitative relationship between two or more variables, linear regression is a linear method in statistics. Simple linear regression analysis, among others, refers to the straightforward linear relationship between two variables [1]. A statistic that expresses the strength of the linear link between two variables is the correlation coefficient. Its value is between -1 and 1. Perfectly negative correlation is represented by a correlation coefficient of 1. One variable's value rising results in the value of another variable falling, and the other way around. A perfect positive correlation is represented by a coefficient of 1. There is no linear relation between variables if the correlation coefficient is zero [2].

2.3.2. Random Forest

A classification and regression technique for ensemble learning is Random Forest. Multiple decision trees are combined, and the result is a class whose output is the average of the output of the class from each tree individually. A random selection of features is chosen to decide the appropriate split for each branch of a decision tree, increasing the model's diversity and lowering the likelihood of

DOI: 10.54254/2754-1169/20/20230199

overfitting. Through the combination of numerous trees, the method lowers the variance of each individual tree, which enhances job performance [3].

2.3.3. Principal Component Analysis

Principle Component Analysis is very useful when there is multidimensional data. Through several steps of calculation, high-dimensional data is converted to low-dimensional data and the data can be explained by those calculated factors clearly and simply. However, one variable may contain of different units, so it is also a hard work to explain the meaning of the factors [4].

3. Results

3.1. Data Visualization

First, each variable is plotted into histograms.

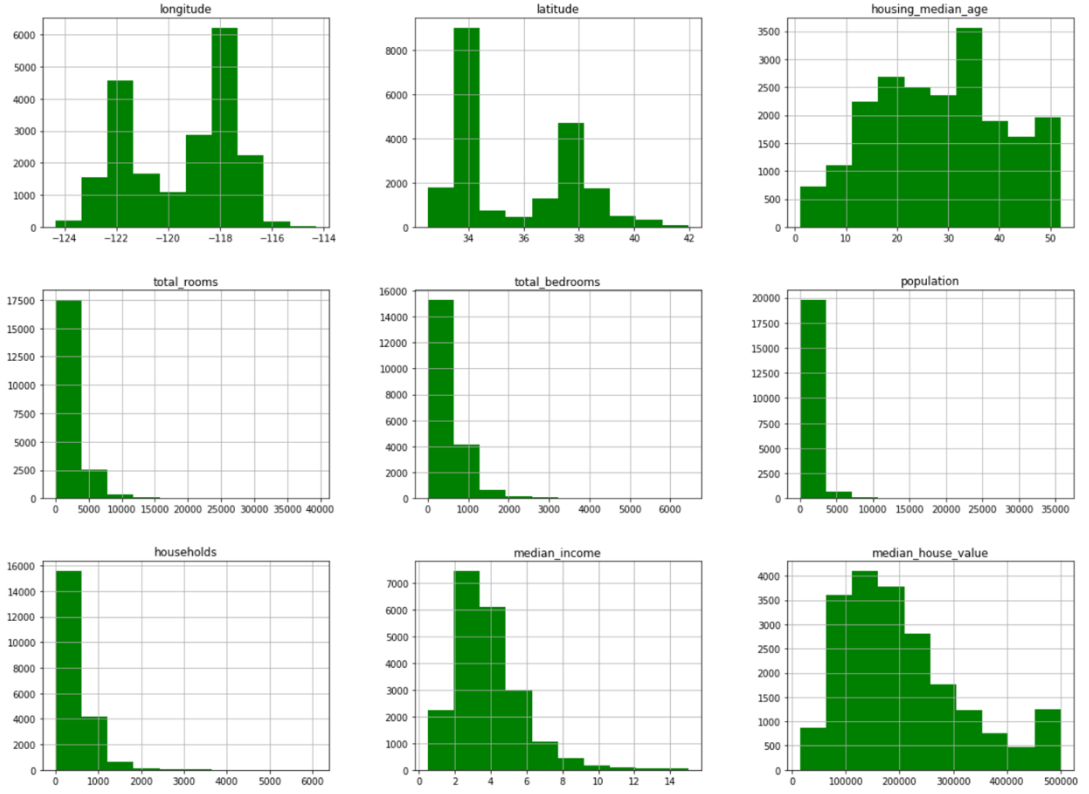

Figure 1: Histograms of each variable (Photo Credit: Original).

From Figure 1, it is clear that most of histograms are skewed to the right, which means the Mean is Greater than the Median. Thus, based on this tendency and the peak of each variable, a rough estimate about median and mode of the data can be obtained.

DOI: 10.54254/2754-1169/20/20230199

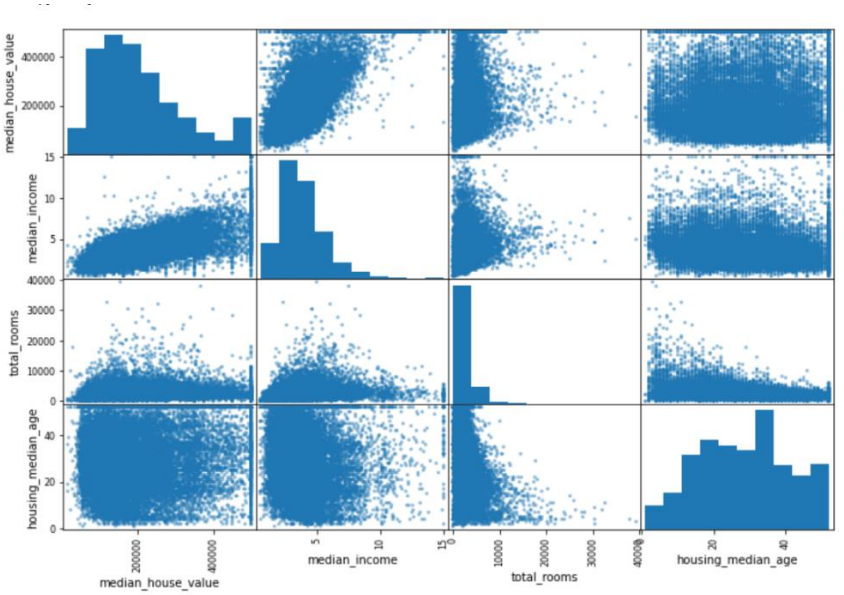

Figure 2: Scattered plots of each variable (Photo Credit: Original).

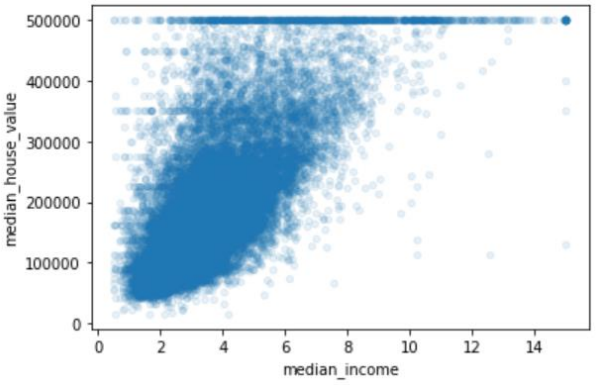

Figure 3: Scattered plot of median income and median housing value (Photo Credit: Original).

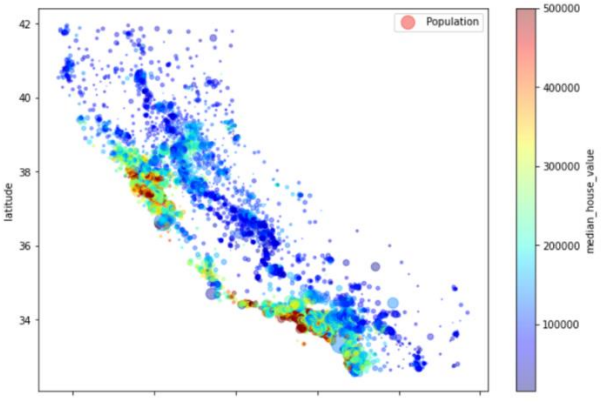

Figure 4: Population's and location's relation with the housing price (Photo Credit: Original).

DOI: 10.54254/2754-1169/20/20230199

Then, using the scatter matrix, Figure 2, to draw the relationship between each variable and the median housing price. By drawing the scatter plot, it can be seen that the variable with the greatest potential to predict the median house price is the median income. By closely looking at Figure 3, there is a clear linear upward tendency between them, and the points are not very separated.

For the Figure 4, the radius of each circle represents the population of each region, and the color represents the price. Therefore, through this picture, it can be known that geographical location (such as the distance to the sea) is closely related to population density.

3.2. Linear Regression Test

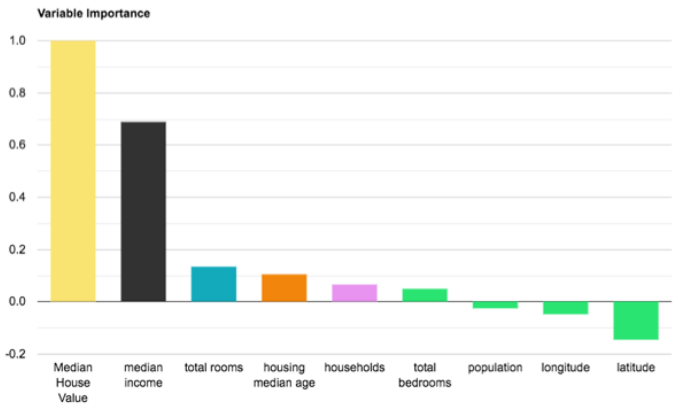

Figure 5: Feature importance (Photo Credit: Original).

Table 2: Correlation coefficient.

Median housing value	1.000000
Median income	0.688075
Total rooms	0.134153
Housing median age	0.105623
Households	0.065843
Total bedrooms	0.049486
Population	-0.024650
Longitude	-0.045967
Latitude	-0.144160

3.2.1. Result Presentation

Finally, by using linear regression to calculate the correlation coefficient of each variable and median housing price (results shown in Table 2), the correlation coefficient ranges from -1 to 1, when the value is close to 1, it represents a strong positive correlation, otherwise, it is a strong negative correlation. The relative importance of each variable is listed in figure 5 according to correlation coefficient. From Table 7, it can be seen that the value of the house is inversely proportional to the number of rooms owned by everyone. At the same time, the lower the proportion of the number of bedrooms to the total number of rooms, the higher the house price. Through these Related coefficients, the prediction of the future Median Housing Value is about 116598.48588.

DOI: 10.54254/2754-1169/20/20230199

3.2.2. Result Analysis

In conclusion, there are some theories to explain why some variables are so crucial. It is evident from the correlation graph between income and housing price that there is a clear positive correlation between the two. Thus, income is the most significant element influencing housing prices, which is understandable. Richer people have higher expectations for their living conditions, which leads them to purchase more expensive homes [5]. Secondly, a larger living space corresponds into a greater price for housing since the house takes more area. Thirdly, the cost of a home increases as construction takes longer. This could be as a result of the locals' increased focus on durability, which can be measured by building time. Fourthly, houses nearer to water sources cost more. This demonstrates that people choose to reside in areas close to rivers or the sea because these areas generally have a better environment, more picturesque surroundings, and quick access to ample water resources. They also relate to house price in terms of latitude and longitude [6]. Fifthly, it makes sense that the interest rate also has a negative impact on housing prices. People are more likely to borrow money to buy a home when interest rates are low, which raises the cost [7]. In addition, there are less limits on the utilization of mortgage loans in the US due to laxer regulation and larger down payment requirements. This raises the possibility of accumulating riches through real estate [8].

3.3. Random Forest

3.3.1. Algorithm Steps

It is unknown which modelling method provides better predictions of the housing price, Although the Random Forest is a powerful method, but it still requires a good understanding of the data. It is still necessary to try different approaches to see which one works best for the specific dataset, cause the actual implementation of the model may vary depending on the specifics of the dataset.

It is better to separate the whole process to six parts to explain. The first part is Exploratory Data Analysis (EDA): Before building the model, it is important to understand the data by doing EDA. For example, to use various visualization techniques to get a sense of the distribution of the target variable, the relationships between the features, and identify any missing values, outliers or other issues that need to be addressed. The second is Data cleaning and pre-processing. This might entail scaling or altering the features as well as adding missing values and eliminating outliers. The third step entails dividing the data into sets for training and testing. It is important to note that the model was trained using the training set, and its performance was assessed using the testing set [9].

To training the model, import the packages from the sklearn library. Then, fitting the model to the training data by passing in the feature matrix and target variable as arguments. The fifth is evaluating its performance on the testing set by using various evaluation metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), etc. For the final part, if the performance of the model is acceptable, it can be deployed for prediction on new data. This involves making predictions on the new data using the trained Random Forest model [3].

3.3.2. Result Analysis

R squared, a measure of the degree of certainty, is 75.7% and it is a metric measuring the goodness-of-fit of a linear regression model. This statistic shows the proportion of the dependent variable's variance that the independent variables account for collectively. This R squared number demonstrates how well the prediction fits the data. Additionally, the submission form has a prediction of the price of housing that was created using Random Forest.

DOI: 10.54254/2754-1169/20/20230199

3.4. Principle Component Analysis

3.4.1. Algorithm Steps

It is challenging and complex to assess every variable in the data set because each one contributes to the house price and there are nine of them. It is particularly challenging to draw conclusions from just the correlation coefficient when two variables are interdependent. Because it converts a set of potentially correlated variables into a set of linearly uncorrelated variables termed principal component by an orthogonal transformation, principle component analysis is a wonderful way to reach an intuitive conclusion about the effectiveness. This method consists of five steps, the first of which is the most crucial because it serves as the foundation for the remaining four.

The first step is to use correlation coefficients instead of primary data and draw the heat map, which is in Figure 6 and Figure 7. Figure 6 shows the correlation efficient before the logarithm operation is done. Figure 7 shows the correlation efficient after the logarithm operation is done. It is intuitive that almost each pairwise variable has a stronger relationship after taken the logarithm.

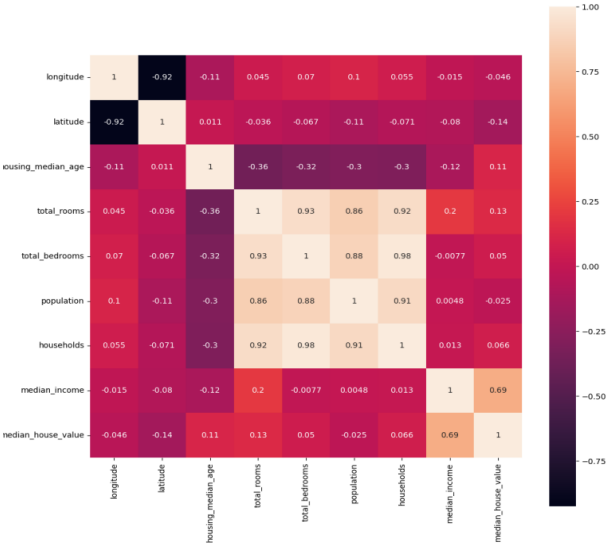

Figure 6: Heat map.

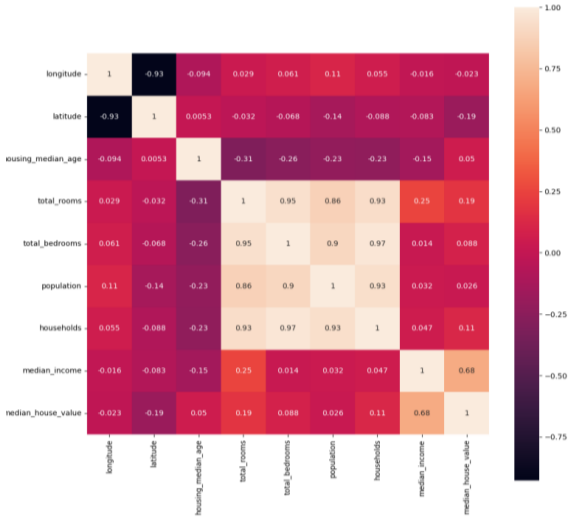

Figure 7: Heat map using logistic data.

DOI: 10.54254/2754-1169/20/20230199

The second step is to calculate eigenvalues and eigenvectors and the partial values are showed in following Table 3 and Table 4.

Table 3: Overview of each variable eig_value.

	Names	Eig_value
0	Longitude	3.911159
1	Latitude	1.902465
2	Housing median age	1.075383
3	Total rooms	0.822217
4	Total bedrooms	0.144516
5	Population	0.076638
6	Households	0.052887
7	Median income	0.014735

Table 4: eig_vector.

	Longitude	Latitude	Housing median age	Total rooms
Longitude	0.076164	-0.701431	-0.055296	0.072459
Latitude	-0.074133	0.701816	0.007810	0.104059
Housing median age	-0.217994	0.015707	-0.095958	-0.885798
Total rooms	0.483619	0.076519	0.095958	-0.119163
Total bedrooms	0.490271	0.059900	-0.116496	-0.063782
Population	0.472615	0.026845	-0.116419	-0.075963
Households	0.491581	0.063477	-0.110803	-0.098269
Median income	0.043729	-0.031760	0.890177	-0.406925

The third step is to draw the scree graph in Table 3 based on eigenvalues, find out a point of rapid descent and determine the number of factors. In the plot, the second point has a sharp decline, but the fourth point still contribute about 1.0 eigenvalues to whole eigenvalues. So, the first four points are selected to be the principle component.

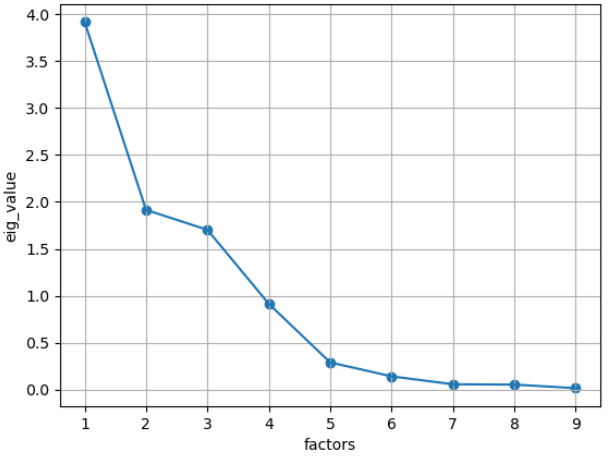

Figure 8: Line Chart of eig_value.

DOI: 10.54254/2754-1169/20/20230199

Table 5: Contribution rate.

1	
Factor 1	3.900349
Factor 2	1.903889
Factor 3	1.106667
Factor 4	0.837814
2	
Factor 1	0.487544
Factor 2	0.237986
Factor 3	0.138333
Factor 4	0.104727
3	
Factor 1	0.487544
Factor 2	0.725530
Factor 3	0.863863
Factor 4	0.968590

The forth step is to calculate the contribution rate, which is shown in Table 5. The result reaches a 96.86% and it shows that the first four factors explain the data set well, which also corroborates the analysis in the third step.

The last step is to do the orthogonal rotation in order to highlight the typical variable of each factor and make it easier to detect the effect of the factor. In Table 6, four factors are associated with several variables and explain the data set well [4].

Table 6: Orthogonal rotation.

	Factor 1	Factor 2	Factor 3	Factor 3
Longitude	0.079621	0.992281	-0.014430	-0.078363
Latitude	-0.078230	-0.993413	-0.060309	-0.027427
House median age	-0.695924	-0.078182	-0.212941	0.677730
Total rooms	0.984890	0.044528	0.064179	-0.143962
Total bedrooms	0.986046	0.073301	-0.084972	-0.111899
Population	0.978488	0.114067	-0.083826	-0.106698
Households	0.988813	0.069553	-0.077473	-0.098609
Median income	-0.069762	0.035482	0.991611	-0.075291

3.4.1. Result Analysis

The heat map shows that median income contributes the most to housing prices. After principle component analysis is taken, whole variables are combined into four factors and the factors have a contribution of 96.86%, which is a successful dimensional reduction. Finally, after the orthogonal rotation, it shows intuitively which variable is contained in a factor. Factor one, also known as the Room Factor, includes the total number of rooms, bedrooms, people, and homes. Factor two contains longitude and latitude, which can be named as Geographical Factor. The Third factor contains median income and can be named as Income Factor. The last factor contains the housing median age and can be named as Housing Age Factor. All in all, all the factors contribute to the housing price and can be divided into different factors, which proves that PCA is suitable and successful for this data set.

DOI: 10.54254/2754-1169/20/20230199

3.5. Discussion

3.5.1. Limitations and Future Directions

Although this paper uses three different methods to analyze the data set and predict the house price, there are still shortcomings. First of all, the data set in this article comes from the 1990s to 2000s, which is a bit far away from now, and the accuracy and reliability of the predictions need to be verified. Second, the data is all from California, and the rest of the United States is not mentioned. At the same time, variables such as the mobility of people, social class, and built factors of different cities are not considered. In future research, more modern data can be obtained and housing prices in different cities can be included in the research scope, and more variables can be added to obtain more accurate predictions.

3.5.2. Suggestion

According to the above research and other corresponding research, the current housing prices are constantly rising. This refers to the strong demand for homes in the U.S. and a record-low supply of housing. The FHFA report pointed out that since September 2011, house prices in the United States have been rising year by year, but the epidemic has exacerbated the rate of increase. The data for the first quarter of 2021 shows that almost every metropolitan area in the United States has seen an increase in housing prices, with the median growth rate of home prices for homes reaching 16.2%, the highest since 1989 [10]. In order to deal with this rise, this article will make several suggestions. First, home buyers need to grasp the laws of real estate as much as possible, especially focusing on changes in national policies and monetary policies. Second, pay close attention to the future planning and positioning of the area where the house is purchased, understand the real estate development situation in that area, and have a clear understanding of its specific real estate price. Third, if buyers are self-occupied, as long as you have enough economic strength, you don't need to pay too much attention to the price, because real estate will always appreciate in the long run. But if it is an investment, buyers need to pay close attention and invest cautiously [11].

4. Conclusion

This article focuses on the relevant factors that affect housing prices in California. The data set is analyzed by three methods (linear regression, random forest, and principle component analysis), and the median house price is predicted, which is 116598.48588. Moreover, according to the R square index (75.7%) of random forest, the predicted value and the data set have a high degree of fitting. The final principle component analysis also confirmed the results of linear regression—the median income is the most important factor affecting housing prices. At the same time, location, housing prices, age of houses, and local housing purchase policies also have a certain impact on housing prices. At the same time, through the fact that house prices are rising, this article also puts forward some suggestions for home buyers, focusing on policies, supporting facilities in the area where houses are purchased, and future economic trends, etc. Finally, make the final purchase decision after careful consideration. Of course, there are some problems in this article, such as the data source being too old, and the geographical restrictions being relatively serious. I hope that future research can conduct more comprehensive and accurate analysis and prediction based on this article.

References

[1] *Su, X., Yan, X., & Tsai, C.-L. (2012). Linear regression. Wiley Interdisciplinary Reviews: Computational Statistics, 4(3), 275–294. https://doi.org/10.1002/wics.1198*

[2] *Taylor, R. (1990). Interpretation of the correlation coefficient: A basic review. Journal of Diagnostic Medical*

DOI: 10.54254/2754-1169/20/20230199

Sonography, 6(1), 35–39. https://doi.org/10.1177/875647939000600106

[3] *Biau, G., & Scornet, E. (2016). A Random Forest Guided Tour. TEST, 25(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7*

[4] *Abdi, H., & Williams, L. J. (2010). Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2(4), 433–459. https://doi.org/10.1002/wics.101*

[5] *Guo, L. (2018, June). Research on the Influencing Factors of American Housing Prices and Its Enlightenment. Research and Enlightenment on Influencing Factors of American Housing Prices--2018 Master's Thesis of "Hebei University". Retrieved February 2, 2023, from https://cdmd.cnki.com.cn/Article/CDMD-10075-1018957687.htm*

[6] *Wang, Y., Wang, S., Li, G., Zhang, H., Jin, L., Su, Y., & Wu, K. (2017). Identifying the determinants of housing prices in China using spatial regression and the geographical detector technique. Applied Geography, 79, 26–36. https://doi.org/10.1016/j.apgeog.2016.12.003*

[7] *Cheng, L. (2012). The Influencing Factors of the Dynamic Changes of American House Prices--An Empirical Study Based on the Perspective of Natural Expectations. World Economic Outlook. Retrieved February 2, 2023, from https://www.zhangqiaokeyan.com/academic-journal-cn_world-economic-outlook_thesis/0201262041127.html*

[8] *Dustmann, C., Fitzenberger, B., & Zimmermann, M. (2021). Housing expenditure and income inequality. The Economic Journal, 132(645), 1709–1736. https://doi.org/10.1093/ej/ueab097*

[9] *Afonso, B., Melo, L., Oliveira, W., Sousa, S., & Berton, L. (2019). Housing prices prediction with a deep learning and Random Forest Ensemble. Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC). Retrieved February 2, 2023, from https://sol.sbc.org.br/index.php/eniac/article/view/9300*

[10] *Li, X., & Zhou, Z. (2021, June 7). The Logic of America's Soaring House Prices. people.cn. Retrieved February 2, 2023, from http://house.people.com.cn/n1/2021/0607/c164220-32124299.html*

[11] *Zhao, T. (2020). Analysis of Factors Affecting Housing Prices Based on Multiple Regression Model. China Business. Retrieved February 2, 2023, from https://www.zhangqiaokeyan.com/academic-journal-cn_china-business_thesis/0201278488564.html*